TRANSATLANTIC WORKING GROUP

Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry[†]

Mark MacCarthy, Georgetown University[1]

February 12, 2020

Contents

Executive Summary

This paper sets out a framework for transparency on the part of the larger digital social media companies in connection with their content moderation activities and the algorithms and data that involve the distribution of problematic content on their systems. It departs from the movement in many countries for content regulation and mandated takedowns, preferring instead to focus on creating a balanced and clear legal structure for disclosure that can help to restore public trust in digital platforms and provide assurances that they are operating in the public interest.

It recommends a tiered system of transparency. Disclosures about content moderation programs and enforcement procedures and transparency reports are aimed at the general public. Disclosures about prioritization, personalization and recommendation algorithms are provided to vetted researchers and regulators. Vetted researchers are also given access to anonymized data for conducting audits in connection with content moderation programs, while personal data and commercially sensitive data are available only for regulators.

This recommended transparency approach could be started through voluntary measures undertaken by the larger social media companies in conjunction with public interest groups and researchers, but

[†] 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/.

its natural home is within a comprehensive system of regulation for the larger social media companies overseen by a government agency.

Transparency is the recommended approach in this paper for several reasons. Openness is an essential element of due process procedures recognized in civil liberties standards, principles for content moderation, international human rights principles, and U.S. administrative law. It is especially important to apply this principle of openness to the larger social media companies, which are the ones to which initially the transparency requirements would apply.

Transparency is also a key element of other accountability measures that have been widely discussed. Those include an independent oversight board that would hear appeals concerning social media content decisions, a special internet court that would use local law (rather than platform community standards) to render expedited judgments on whether certain content should be removed from platforms, and social media councils to oversee content moderation practices by the major platforms. The recommendations in this paper are likely to accommodate the information needs required by these external reviewing institutions.

Improved transparency also enables the forces of consumer choice to do their work, empowering platform users to protect themselves and to bring the pressure of public opinion to bear on social media companies.

Better transparency might also create an interactive public policy dialogue that could gradually scale up regulations as needed, improving their structure and stringency on the basis of feedback. This process of improvement could apply to the transparency measures themselves or to broader mandates, such as for a duty of care. The cycle would be to issue a guideline, implement it, assess it, retrofit it, enrich it and start again.

Finally, transparency does not raise the free expression issues that bedevil mandated requirements for removal of problematic material. In the United States, First Amendment jurisprudence is uniformly hostile toward content-based regulation and likely prohibits the government from directly requiring removal of legal content. In Europe, the protection of free expression is one of the fundamental human rights enshrined in the various charters that bind the European countries and is also embodied in national legislation. The focus on transparency measures might provide an effective approach to avoiding these obstacles.

The paper begins with a survey of current law, practice and proposals for reform in the area of transparency. Some laws explicitly require social media companies to issue transparency reports describing how their content moderation programs operate. In addition, the companies have voluntarily disclosed information about their programs and shared some information with outside researchers to assess the performance of these programs. Moreover, outside researchers using publicly available data can often discern much about the operation of social media algorithms used to prioritize, personalize, and recommend social media content.

While current platform practices provide real transparency in some regard, the overall insight into platform operations and decision making is limited. Moreover, current platform practices and legal requirements seem unlikely to move the platforms closer to socially desirable levels of disclosure, at least in the short term.

The report surveys various proposals for transparency reform from interest groups, academics, and policy makers seeking to improve the public transparency reports, the information provided to regulators and the data available to vetted researchers.

The heart of the paper is a series of recommendations to improve transparency. They can be summarized as follows and are illustrated in Table 1:

1. Continued and improved public disclosure of the operation of platform content moderation programs, including:
    a. Content rules in terms of service or community standards;
    b. Enforcement techniques such as deleting, demoting or delaying content;
    c. Procedures for the public to complain about possible rules violations;
    d. Procedures for platforms to explain their decisions to affected parties; and
    e. Procedures for individual appeals in connection with enforcement actions.
2. Continued and enhanced reports to government agencies and to the public with aggregate statistics accurately reflecting the operation of the content moderation programs.
3. Technical terms of reference of algorithms used in content moderation, prioritization and recommendation.
4. Greatly improved access to platform data for qualified independent researchers and regulators. Access to information must be in a form and quantity to permit regular and ongoing audits of these platform operations to verify that they are operating as described and intended and should include data relevant to:
    a. the operation of content moderation programs;
    b. sponsorship of political advertisement; and
    c. content-ordering techniques, including recommendation and prioritization algorithms.

The proposals in this working paper are designed to further the public's interest in the transparent operation of digital social media platforms with the aim of ensuring that the platforms' operation furthers the twin interests in effective content moderation and a robust environment for free expression on crucial matters of public importance.

	Public	Vetted Researcher	Regulator
Information Type			
Content Moderation Program			
Content Rules	Yes	Yes	Yes
Enforcement Procedures	Yes	Yes	Yes
Complaint Process	Yes	Yes	Yes
Explanations	No (users)	No	No
Appeal Rights	Yes	Yes	Yes
Reports			
Content Moderation	Yes	Yes	Yes
Algorithms (Technical Description)			
Content Moderation	No	Yes	Yes
Prioritization	No	Yes	Yes
Recommendation	No	Yes	Yes
Data			
Content Moderation Program	Yes	Yes	Yes
Political Ads	Yes	Yes	Yes
Content-Ordering Techniques	No	Yes	Yes
Commercially Sensitive/Personal	No	No	Yes

Table 1. Disclosure Recommendations by Audience and Information Type

Introduction

Global concern about the use of digital social media platforms for hate speech, terrorist material and disinformation campaigns has prompted governments to pass or consider legislation that requires platforms to remove certain kinds of speech. In 2017, Germany adopted its network enforcement law (NetzDG), which requires platforms to remove content that is illegal under a wide variety of local law.[2] In 2019, the French Assembly approved a measure modelled on the German NetzDG law requiring social media networks to remove hate speech within 24 hours.[3] In 2019, the European Parliament backed a terrorist content proposal that mandates removal of terrorist material within one hour of notification.[4] A similar measure to mandate content removal is pending in the United Kingdom, which has proposed a duty of care that would require platforms to take down certain harmful content.[5] In the wake of the widespread online distribution of the Christchurch video, Australia has adopted a law that would outlaw the sharing of violent abhorrent material.[6] Singapore's Online Protection From Online Falsehoods and Manipulation Act, which went into effect on October 2, 2019, bars the communication of "false statements of fact" and provides extra penalties if this is done on digital platforms through inauthentic accounts.[7]

This working paper takes a different approach. It agrees with former U.S. Supreme Court Justice Louis Brandeis that sunlight is the best disinfectant and calls for policy makers to explore various

transparency measures for digital social media platforms. A balanced and clear legal structure for disclosure can help to restore public trust in these platforms and provide assurances that they are operating in the public interest.[8]

This approach requires enabling legislation to mandate certain disclosures and to establish a regulatory agency to supervise these legally mandated disclosures. The agency should have full authority to mandate additional disclosures as needed over time. The paper recommends that the transparency regime for addressing hate speech, disinformation campaigns, and terrorist material should be part of a larger regulatory structure to ensure that digital platforms are operating in the public interest.

Some may question the need for a regulatory agency with authority to supervise the larger digital social media platforms. But these platforms meet the requirements for special regulatory treatment that have motivated the creation of such agencies for the communications and financial services industry: they are central to our public life and competition has persistently failed to ensure their operation in the public interest. Digital social media platforms have installed themselves at the heart of our societies, in the cauldron of public opinion, sitting right next to the traditional communications media. Moreover, they convey and indeed amplify content that reflects some of the worst aspects of our societies, namely, hate speech, disinformation campaigns, terrorist material, child exploitation images, harassment and bullying. And typical business incentives are unlikely to remedy this content disorder on their own. For these reasons, a comprehensive regulatory response is needed.[9]

The transparency measures recommended in this working paper are an essential element in this regulatory structure. These might not be the only measures needed. The information uncovered through mandated disclosures will contribute to the ongoing policy conversation on the best ways to structure balanced, flexible regulations. This might lead ultimately to some forms of content regulation, or a duty of care, crafted to accommodate the demands of an open system of free expression. While this paper leaves that possibility open, it does not recommend measures beyond transparency.

Additional transparency measures might be needed to deal with broader problems of subtle dark patterns and targeting techniques that threaten the integrity of consumer interactions with social media platforms and expose users to abuse and manipulation.[10] They might be required to expose discriminatory practices in advertising, where the targeting criteria for campaigns in connection with housing, employment and credit granting might be disproportionately adverse to members of protected classes.

But these concerns and additional regulatory measures to address them are outside the scope of this working paper, which is focused exclusively on transparency measures that might respond to the problems of hate speech, terrorist material, and disinformation campaigns.

Transparency measures are needed to reveal the operation of algorithms on digital social media platforms in order to address content moderation concerns, but this paper does not address the related question of the extent to which algorithms can successfully identify harmful material, nor the question of whether recommendation and prioritization algorithms can increase, intentionally or not, the distribution and salience of harmful material. This paper does assume, however, that platforms need to disclose enough information about their algorithms and the data used to train them so that regulators and researchers can form judgments about these vital questions, which should not be left to the sole discretion and judgment of the platforms themselves. This paper will focus on which data and features of algorithms social media platforms should disclose and to whom in order to allow accountability assessments to be made by regulators and independent researchers concerning the

operation of platform algorithms related to content moderation and the distribution of problematic content.

The form of regulation envisaged in this paper calls for digital social media platforms to reveal what they are doing in content moderation and in the content-ordering algorithms that might exacerbate the distribution of harmful material. But it does not mandate any particular moderation practices. Platforms are free to moderate whatever content they feel is appropriate; their only obligation under the recommendations in this paper is to tell the user, the regulatory agency, and the public what their policies are and how these policies are enforced. The diversity of content moderation practices in the current social media world would persist under this recommendation.

However, this freedom of choice for the platforms creates obligations once it is exercised. Under the recommendations in this paper, digital social media platforms are free to make promises to the public concerning their content moderation practices, or not, as they see fit. But they are not free to make promises to their users that they do not keep. The supervising regulatory agency would be authorized to enforce these promises as well as any disclosure obligations to ensure that the public, the regulators and researchers have sufficient information about how platforms' moderation practices and content-ordering techniques might exacerbate the distribution of problematic content.[11]

This combination of individual platform choice backstopped with regulatory control might be further refined. The regulatory framework for financial broker-dealers in the United States might be a model for an additional step, moving from company-specific decision making toward collective regulation by and of the industry itself.[12] In this model, digital social media platforms would not be free to choose whatever content moderation practices they wanted. Rather, these practices would be set by an industry association and would be binding on all members of the association. This collective industry organization would enforce the rules, with power to investigate complaints, inspect business operations and punish offenders with fines, suspension and ultimately expulsion.

This model of collective industry regulation has the great merit, from a content regulation point of view, that no government body sets content rules for the industry. It is a matter for the industry itself to determine, not for regulatory determination. But at the end of the day, it is government that compels obedience to these industry-set rules through a requirement that all members of the industry be licensed or approved by a professional trade association.

This paper does not conclude that such a collective regulatory structure is needed or desirable. The requirement for licensing by a professional trade association creates a potential for industry self-censorship that should give pause to all who care about free expression, and might raise First Amendment issues in the U.S. But it is a possible evolutionary path for digital social media platforms that stops short of overt content regulation by a government agency. The mandated disclosures recommended in this report might help to clarify whether movement in the direction of collective self-regulation is needed.

While this paper recommends a comprehensive regulatory structure and mandated disclosures as a part of that structure, it does not suggest waiting for policy makers to perfect legislative measures. The platforms themselves, often in conjunction with policy makers and outsider researchers, have already adopted transparency as a governance mechanism that can increase public trust in the proper operation of their systems. Much is being done on a voluntary basis, and more could be done without the need for further government mandates.

The voluntary Christchurch call, for instance, which has been signed by numerous governments and the major platform companies, commits platforms to a range of measures to combat terrorist and violent extremist content including "increasing transparency around the removal and detection of content, and reviewing how companies' algorithms direct users to violent extremist content."[13] The Social Science One initiative – described later – is a workable, though troubled, framework for voluntary efforts in this area, as is the agreement between tech companies and the European Commission in connection with disinformation campaigns. Many of this paper's specific recommendations can be incorporated into these ongoing efforts.

The advantages of transparency have often been noted. Transparency is an essential element of recognized due process procedures, including the civil liberties standard called the Manila Principles,[14] the Santa Clara Principles for content moderation,[15] international human rights principles,[16] and the due process tradition in U.S. administrative law that typically provides individuals meaningful opportunities to challenge adverse decisions.[17]

Transparency is also a key element of some external accountability measures that have been called for by several commentators. Facebook is working with outside groups to establish an independent oversight board that could hear appeals from content decisions made by moderators working for Facebook.[18] Others recommend a special internet court that would use local law (rather than platform community standards) to render expedited judgments on whether certain content should be removed from platforms.[19] Still others want social media councils that would address and oversee content moderation practices by the major platforms.[20]

Clearly, substantial information disclosure is needed to make these accountability mechanisms effective. While the recommendations in this paper are likely to accommodate the information needs required by these external reviewing institutions, it is not part of this paper's mission to recommend or discuss the need for external reviews of content moderation decisions. The paper takes a small step toward accountability measures by recommending that platforms allow those whose request for the removal of content is denied to ask for a review, in parallel with their current practice of allowing users whose content has been removed to ask for a second look.[21] To accommodate both types of review, this paper recommends that platforms provide a reference to the specific community standard that justifies a removal action or permits certain content to remain visible.

Transparency rules are consistent with the disclosure philosophy in investor protection laws that require public companies to provide disclosures about their financial condition, operating results, management compensation, and other areas of their business, and prohibit deceit and misrepresentation in the sale of securities.[22] Transparency is also at the core of consumer protection laws that forbid companies from engaging in unfair or deceptive acts or practices in connection with the sale of goods or services to the public.[23]

One objective of transparency rules is to enable the forces of consumer choice to do their work. If consumers and investors have accurate information, they are empowered to purchase only the products, services, and securities they find attractive. Transparency rules for digital platforms can serve the same objective of empowering platform users to protect themselves. They do this by requiring platforms to detail the content rules and enforcement procedures they use and to publish regular reports on the operation of their content moderation systems, and by allowing external access to platform data for researchers and regulators to conduct audits to describe to the public how the systems work and to enable an assessment of whether that operation is in the public interest.

An additional governance function of transparency rules is to bring the pressure of public opinion to bear on digital platform operations. Companies are often moved to change when their conduct violates well-entrenched social norms, even when the conduct itself is not illegal. Even when rebroadcasting hate speech or terrorist material is legal under local law, for instance, companies whose policies permit that face severe public pressure not to air such material.[24]

Sometimes social platforms do not want to know whether their platform moderation enforcement procedures or personalization, ordering and recommendation algorithms are having certain effects either within their own platform or in the external world.[25] However, it might be desirable for platforms, even from their own point of view, to do some of this work themselves. For instance, they might find it a wise investment in compliance to conduct disparate impact assessments of their advertising practices to see whether their facially neutral algorithms produce disproportionate adverse impacts on protected classes.[26] They might also find it useful to assess whether their content moderation practices, while not overtly partisan, nevertheless produce outcomes that favor one political perspective over others.[27] It would be possible for them to hire external auditors to check their systems for these effects in a system akin to that of using a financial auditor.

But a big advantage of transparency requirements is that this moves information out from the platforms to the public so that these studies, audits and assessments can be performed independently. In this way, even if the companies have an interest in preserving ignorance, or simply have no rational basis to find out certain things on their own, researchers outside the company have the resources they need to fill the gap and provide the public and regulators with these studies.

Release of information to the public, to experts working for government agencies and to independent researchers working for think tanks, civil society organizations or universities might also create an interactive public policy dialogue that could gradually scale up regulations and improve their structure and stringency on the basis of feedback. This process could apply to the transparency measures themselves or to broader mandates such as for a duty of care. The cycle would be to issue a guideline, implement it, assess it, retrofit it, enrich it, and start again.

Identifying the purposes of transparency requirements helps to clarify a key element of legislation to implement these requirements. These purposes also serve as the objectives the legislation is seeking to achieve and that govern the activity of the regulatory agency established to interpret and enforce these requirements. At the most general level, transparency serves to reveal to the public the content rules a social media company has developed and how well it is living up to those rules. The disclosures also allow researchers and the public to determine the effects that the operation of the social media company is having on a range of social variables, including the prevention of the spread of harmful speech, the preservation and promotion of freedom of expression, and the impact on political processes. Legislation addressing the transparency of content moderation practices of platforms does not raise the free expression issues that bedevil mandated requirements for removal of problematic material. In the United States, First Amendment jurisprudence is uniformly hostile toward content-based regulation and generally prohibits the government from directly requiring removal of legal content. In Europe, the protection of free expression is one of the fundamental human rights enshrined in the various charters that bind the European countries together and is also embodied in national legislation. The focus on transparency measures might provide an effective approach to avoiding these obstacles.[28]

Takedown approaches face another difficulty, namely the extent to which takedowns are national or global in scope. On the one hand, local takedowns seem appropriate for content rules that might vary by jurisdiction. On the other hand, it does little good to block genuinely dangerous content only in a

single jurisdiction. Recent European court decisions send a mixed message on whether European takedown rules are regional or global in scope. European governments are permitted to mandate worldwide takedowns for defamation.[29] But under current EU law, removal of privacy-violating material is limited to Europe.[30]

In contrast, the transparency approach recommend in this paper has global benefits: what is released to the public anywhere is generally available everywhere. Networks of regulators in different countries could assure that information shared with one national regulator is also available to regulators in other countries with a similar mission.

Still, the transparency approach does not entirely escape jurisdictional issues. Many of the social networks are global companies with operations that cross many jurisdictions. When a jurisdiction requires companies to disclose information about its operations, does this apply to operations outside its own jurisdiction? Rather than remaining silent on this jurisdictional question and leaving the decision up to later court interpretation, legislators should specify the jurisdictional reach of the transparency requirements recommended in this paper. But the issue of which jurisdiction is more appropriate is beyond the scope of this paper.

Other issues will need to be addressed that are also outside this paper's scope.[31] One concerns the geopolitical implications of social media transparency requirements. China insists that any social media company doing business in China must accept its local content laws, with the result that many U.S. companies choose not to do business in China. Similarly, local laws in the U.S. and Europe apply to Chinese companies doing business there, and this would apply to any new transparency requirements. A foreign company would not be able to have a secret algorithm that blocks content they find objectionable, but which is hidden from users and from the public. Chinese companies must comply with these laws if they want to do business in these jurisdictions. As a result of increasing divergence of local laws, and the rise of "techno-nationalism," which treats technology as intrinsically connected to national security issues, the integration of major economic actors has slowed and may even be reversed in the years to come.[32] Transparency rules will inevitably be part of any such "decoupling" of the world's major economies. But the implications of this are outside the scope of this paper.

A further issue concerns the interface of transparency rules with law enforcement and national security concerns. Some of the information social media companies have to provide to the public, to regulators and to vetted researchers under new transparency rules will have value for law enforcement and national security purposes. Data that would enable audits, for instance, might allow identification of specific individuals or types of individuals and so also allow the construction of profiles of social media users that could be compared with or combined with data on potential terrorist suspects or criminal actors. Should government agencies be allowed to use this data for these purposes? If so, under what due process protections? This complex and controversial issue requires balancing the interests of users to be protected from government surveillance with the needs of national security and law enforcement to fulfil their vital missions. While any new transparency requirements will have to contain a balanced resolution of this issue, it is beyond the scope of this paper.

In addition, transparency of government action in connection with social media platforms is crucial. Years ago, platforms initiated their transparency reports as a way to let the world know the extent of government efforts to affect content on their systems. This is still a matter of crucial public concern. This paper approaches it through requirements for transparency on the part of platforms, rather than additional disclosures by government. If companies are clear about their standards and practices for removal of content that will to some extent reveal government activities. But addressing additional disclosures by government raises questions about the extent of access to government activities through

various open government laws and the extent to which such activities need to remain secret to protect important security and law enforcement activities. The right balance of these conflicting objectives is beyond the scope of this paper.

This working paper proceeds as follows. The next section reviews current law, practices and proposals for reform in connection with transparency. This is a snapshot of the status quo, which, of course, will change perhaps rapidly over the coming months and years. But it provides a baseline from which to consider the improvements that might be necessary. It is organized according to whether the disclosures are directed to the public, to a regulatory agency, or to independent researchers.

Following this background, the paper makes its recommendations for disclosures, which are structured in several levels. The first level is the information that should be made available directly to users so that they might better understand the content moderation process on the platforms they use and take advantage of any complaint or redress mechanisms the platforms provide. The second level is the information that should be in the public reports that the platforms are issuing today. The third level is the information about the operation of the platforms that should be released to regulators and researchers to enable audits of content moderation systems, political advertising, and the content-ordering algorithms that can sometimes exacerbate the distribution of harmful content. Within the third level, it is crucial to distinguish between information that can go to the general public in a form that can be used by any independent researcher and information that is available only to regulators and approved independent researchers.

Current Law, Practice and Proposals for Reform

Internet platform companies operate across enormous swaths of society, facilitating global access to social communications, financial transactions, and information. While billions of people rely on these services daily, little is known publicly about the ways in which these companies operate. This section will examine differing transparency requirements and practices as they currently exist and will outline current proposals to increase transparency. The section will first look at law, practice, and proposals for disclosures to the public and to government agencies, and then will discuss disclosures to academics and researchers.

Disclosures to the Public and Regulatory Agencies

i. Current law in connection with disclosures to the public

Few laws currently require digital social media platforms to make active disclosures to the public or to government agencies. There are currently no U.S. federal laws that mandate disclosures on content policy; indeed, Section 230 of the Communications Decency Act (47 U.S.C. 230) gives online service providers broad latitude to make decisions on how to handle content, and it contains no requirement for disclosure of content moderation practices.

Some states have taken steps to require platforms or online participants to provide greater transparency about their actions. For instance, California recently enacted laws requiring political advertisers[33] and bot operators[34] to disclose information to the public about their activities. Other laws, such as the California Consumer Privacy Act, require transparency about data collection and use.[35] In Europe, the General Data Protection Regulation (GDPR) has a similar requirement for firms that collect personal information to disclose that fact to the subject of the information.[36] These data

protection and consumer privacy disclosure rules can complement the transparency measures called for in this report related to the operation of content moderation policies and procedures.

Outside the United States, other countries have experimented with mandated disclosures. In Germany, the Network Enforcement Act (NetzDG) requires social media companies with two million or more registered users in Germany that receive over 100 complaints about online content per year to submit semiannual reports about how the company handles complaints.[37]

These reports must include information on the actions taken by the platform to remove illegal content, descriptions of how to submit complaints and criteria for handling those complaints, a tally of those complaints and how they were handled, personnel and training metrics for moderators, whether the platform consulted outside organizations when making takedown decisions, and other information about removal statistics and timing.[38] Under NetzDG, social networks must also provide users with open, transparent guidelines for how to submit challenges and file complaints.[39]

The German Office of Justice reviews these public reports and is authorized to issue fines for failure to report enforcement activity adequately and completely.[40] In July 2019, this office fined Facebook for underreporting the number of complaints under NetzDG.[41]

In the United States, several states require disclosures of political ads on digital social media platforms.[42] But these disclosure obligations fall on the political advertisers, not the platform. Political advertisers often fail to follow the requirements.[43]

ii. Current practice in connection with disclosures to the public

Most major platforms publish their community standards for the public to see and evaluate.[44] Some platforms, in addition, publish their enforcement guidelines, which allow the public to see how these general rules are interpreted and applied in particular cases.[45]

In addition to legal requirements to disclose, many online platforms provide voluntary reports in connection with their enforcement of their community standards. These voluntary reports outline much of the same information as required by German law, including overall volume of content reported and removed, as well as information on appeals.[46]

Platforms also sometimes share limited information on an ad hoc basis. When Twitter discovered a coordinated misinformation campaign by the Chinese government targeting protestors in Hong Kong, it shared its finding with Facebook and then made the datasets public.[47] Twitter first announced the action and the bad actors and how the actions violated policies. The platform provided examples of content violating policies[48] and explained how and why it would be updating its advertising policies, including changes in defining certain categories of actors online.[49]

Many platforms, including Twitter[50] and Google (including YouTube),[51] provide public archives of political advertisements. Facebook offers disclosures of political ad sponsors, authentication of political ad sponsors and availability of political ads for research.[52]

Several digital social media platforms, including Google, Facebook and Twitter, signed a voluntary agreement with the European Commission on disinformation, which commits the platforms to disclosures of political ads and issue ads, identifying automated bots, prioritizing authentic information, and not discouraging good faith research into disinformation.[53] The agreement also

requires the platforms to file regular reports with the Commission on their compliance with this voluntary code, which are then reviewed and published on the Commission website.[54]

Platforms including Facebook, Microsoft, Twitter and YouTube also have established and manage a program of knowledge-sharing, technical collaboration, and shared research in connection with terrorist content.[55] This Global Internet Forum to Counter Terrorism (GIFCT) issues a regular transparency report on its work against terrorism.[56]

iii. Proposals for reform of disclosures to the public

A report to the French government suggested a tiered approach to disclosures, with substantial information available to users to help them understand more fully the operation of the systems they use; greater transparency for experts working for government, who can be expected to understand the detailed terms of reference that platforms might release to describe the operation of their systems; and access to data for researchers to conduct studies. An important element of the French proposal is that access to a platform's operational data should be compliant with the GDPR regulation. To the extent that such data includes protected personal information, it would be controlled and made available only to vetted researchers, not to the general public.[57]

In connection with the disclosures under NetzDG, some have expressed concerns that the NetzDG reporting requirements do not mandate a particular format, making cross-platform comparisons of data difficult. Further, these critics say, because platforms are only required to produce aggregate data, the reports do not provide any information about the handling of individual cases, which creates challenges when trying to determine the adequacy or fairness of platform actions. These critics argue that requiring a standard format and additional information on accuracy in individual cases would increase the usefulness of these reports to the public and regulators.[58] These changes may also require privileged access for vetted researchers because some of this content will, by definition, be illegal to publish under German Law.

In the United States, members of the Senate and House of Representatives have introduced legislation to require disclosure of information in connection with political advertising on digital social media platforms. This legislation, the Honest Ads Act, mirrors current laws that require disclosures concerning political ads that air on radio and television. It generally requires information on the sponsor of the ad and would require platforms to maintain public records of political ads.[59]

In addition, Senator Dianne Feinstein has introduced legislation similar to the California law requiring identification of bots. It goes beyond the California law in banning the use of bots in connection with political campaigns and political advertising.[60]

The Institute for Strategic Dialogue (ISD) has suggested several improvements in the area of disclosure of political ads, and in connection with the disclosure of complaints and redress.[61]

The UK White Paper on Online Harms suggests a number of transparency measures aimed at improving public understanding of the operation of content moderation systems, including empowering a regulator to require public annual transparency reports from platforms "outlining the prevalence of harmful content on their platforms and what counter measures they are taking to address these." The UK also recommends that the regulator "have powers to require additional information, including about the impact of algorithms in selecting content for users and to ensure that companies proactively report on both emerging and known harms."[62]

In April 2019, Facebook's Data Transparency Advisory Group (DTAG), a group of independent researchers, released a report assessing Facebook's methods of measuring and reporting on its policies for enforcing its community standards.[63] The report recommended a wide range of improvements in how Facebook should report its enforcement activity.

Though these Facebook transparency reports include quite a few quantitative metrics, they do not provide a qualitative report of enforcement actions by particular types of content or how decisions were made. Similarly, these reports provide raw appeals numbers (i.e., total actions appealed and total pieces of content restored), and requests for legal process, but do not discuss how the process works.[64] The DTAG group notes that these metrics obscure some types of information that would be useful to further understanding and examining moderation practices:

> Qualitative reporting: Transparency reports should include the types of information and examples of takedowns and other adverse actions. For instance, some additional detail about the types of content removed under the category of "removed pornography" would be helpful to enable further discussion about the criteria and removal processes.
>
> Discussion of True and False Negatives: Current transparency reporting focuses on two types of action: removals and appeals. This gives a sense of (1) how much content was removed, and (2) how many removals were later reversed or upheld, that is, it gives true and false positives. To gain a full sense of moderation practices and error rates, reporting should also include how many pieces of information were initially flagged or suggested for removal that were then not removed, that is, true negatives, and attempt to determine how much content is slipping through the cracks and is never identified, that is, the false negatives, the unknown unknowns.

Many of these suggested reforms from these different organizations form the basis for recommendations for improvement in public reporting that are further examined in the next section.

Access to Information for Researchers

i. Current law on access to information

There is no current U.S. law that empowers researchers to access social media data or compels platforms to provide that data to third parties. On the contrary, U.S. criminal law has been used to deter researchers who attempt to obtain information from social media sites by "scraping" – automatically downloading – the sites for information.[65] For instance, the Computer Fraud and Abuse Act has been interpreted to allow websites and platforms to bar outsiders, including researchers, from collecting information that is publicly available on their sites by stating that such scraping is prohibited under the terms and conditions for user access to the website.[66] While a recent court decision has changed the legal landscape in the United States, potentially allowing researchers to scrape sites that do not use technical means to prevent access to publicly available data regardless of the terms of service,[67] there has not been a concerted push to enact a law proactively granting researchers access to platform data. Even without legal powers to prohibit scraping, platforms may still have the technical ability to prevent scraping in practice.

ii. Current practice on access to information

The platform information currently available to the public can allow researchers to uncover important aspects of the operation of various algorithms. For instance, Upturn, a Washington, D.C.-based research organization, was able to examine the advertisement-targeting techniques used by Facebook in order to demonstrate that the platform might be violating U.S. law by creating discriminatory effects in housing advertising.[68]

Several platforms voluntarily provide information to academics for research purposes. In April 2018, Facebook announced Social Science One (SSO), a collaboration between Facebook and a group of independent researchers to use Facebook data to "address societal issues."[69] Facebook made several datasets available to researchers, including information on election advertisements and engagement data.[70]

Proposals to SSO are reviewed by an independent academic panel, which makes recommendations for funding. The first batch of program grants was awarded in April 2019.[71] However, many of the researchers granted awards have not been given access to information (specifically, information on "URL shares," a particular metric of engagement)[72] they were promised, which has prompted foundation funders to contemplate withdrawing financial support from the program.[73] Facebook has said that it cannot provide the information because of privacy issues;[74] specifically, the project initially anonymized data using k-anonymity (a system that removes identifiers until each entry is identical to k other entries, where k is a measure of how hard it would be to reidentify a given user).[75] Researchers and others urged Facebook to instead rely on differential privacy.[76]

In addition to Social Science One, Facebook invites select researchers to work alongside its employees on complex topics such as machine learning and privacy.[77] However, these academics do not necessarily work on key aspects of platform governance, such as moderation or community standards development, and it is not clear that the researchers can publish raw data or provide insights not approved by Facebook as part of their work product.

Facebook has taken some steps to provide researchers with access not just to platform data but to decision-making processes and content decisions. For instance, it has allowed St. John's University professor Kate Klonick to witness and write on the development of the Facebook Oversight Board.[78]

iii. Proposals for additional access to information

The Knight Institute on the First Amendment at Columbia University suggested that Facebook should allow even more access to data for journalists and independent researchers than would be permitted under Social Science One.[79] Mozilla has suggested substantial improvements in the structure of archives of political ads provided by platforms.[80] The recent report to the French government suggested that researchers vetted by a government regulator should be given unfettered access to social media data to conduct accountability analyses.[81] ISD has proposed additional access to information concerning certain platform algorithms to allow audits of the recommendation and prioritization functions.[82] The New American Foundation has called for greater transparency in connection with algorithmic structuring of social media content.[83]

Recommendations

The previous sections have surveyed the landscape with respect to current platform transparency practices, the current legal framework for governing these practices in Europe and the United States, and leading proposals for reform. They investigated transparency along several dimensions:

- the operation of platform content moderation programs
- platform prioritization and recommendation algorithms
- information related to political and issue-oriented advertising
- access to platform information for independent researchers and researchers with government agencies seeking to audit the operation of these programs.

The previous section found that while current platform practices provide real transparency in some regards, the overall levels of insight into platform operations and decision making is limited. Moreover, current platform practices and legal requirements seem unlikely to move the platforms closer to socially desirable levels of disclosure, at least in the near term.

A key problem is a persistent trust gap with policy makers, which undermines the credibility of otherwise positive industry initiatives, such as public transparency reports and the public availability of advertisement libraries for researchers. This leads policy makers to enact or propose strong content-based interventionist measures, such as NetzDG, which could begin to undermine the promise of social media companies as platforms for robust and open discussion of public issues.

Strong transparency practices and requirements can provide the public and policy makers with assurances that platforms have in place policies and procedures reasonably designed to address the challenges of hate speech, disinformation campaigns, and terrorist material. They can also focus public discussion on improvements that policy makers can develop cooperatively with platforms to stay ahead of the evolving threats in this area while continuing to respect the vital platform role as exemplars of open discussion.

This section summarizes the paper's transparency recommendations for policy makers and industry. These focus on:

1. Continued and improved public disclosure of the operation of platform content moderation programs, including:
    a. Content rules in terms of service or community standards;
    b. Enforcement techniques such as deleting, demoting, or delaying content;
    c. Procedures for the public to complain about possible rules violations;
    d. How platforms explain their decisions to affected parties; and
    e. Procedures for individual appeals in connection with enforcement actions.
2. Continued and enhanced reports to government agencies and to the public with aggregate statistics accurately reflecting the operation of the content moderation programs.
3. Technical terms of reference for algorithms used in content moderation, prioritization, and recommendation.

4. Greatly improved access to platform data for qualified independent researchers and regulators. Access to information must be in a form and quantity to permit regular and ongoing audits of these platform operations to verify that they are operating as described and intended and should include data relevant to:

    a. the operation of content moderation programs;

    b. sponsorship of political advertisement; and

    c. content-ordering techniques, including recommendation and prioritization algorithms.

The following sections explore these recommendations in more detail.

A fundamental assumption is that disclosures will be more effective as a governance mechanism if supervised by a government agency with comprehensive regulatory oversight of digital platforms. Several commentators have suggested the establishment of such a government agency with responsibilities to promote competition, protect consumers, enforce privacy rules, and oversee content moderation programs.[84] Disclosure requirements fit naturally within this regulatory scheme.

There are strong arguments that platforms should disclose key elements of the operation of their content moderation programs. In particular, requirements for platforms to say what they do and then do what they say in connection with these programs are needed to allow consumers to make informed choices about using digital platform services.[85] The recommended disclosures in this section will be most effective if they are part of a more comprehensive regulatory framework.

The specific recommendations below are not the final word on transparency measures. They are derived from the recommendations from groups that have reviewed current platform disclosure practices, including the European Commission,[86] the Institute for Strategic Dialogue,[87] and Data Transparency Advisory Group.[88] They also benefit from the due process measures espoused in the Santa Clara Principles.[89] A key benefit of ongoing supervisory efforts by regulators is that they allow the evolution of disclosures to fit the changes in platform technology, threats, and standards that will undoubtedly occur over time. Maintaining regulatory flexibility will allow continued development of technical and platform tools and further essential innovation. Regulators should work with platforms to modify the initial required disclosures to respond to changes in platform policy and infrastructure.

A key element of the recommendations is tiered access, which is needed to accommodate the privacy of platform users and the interests of platform companies in preserving the confidentiality of commercially sensitive information that should not be released generally to the public, but which might be crucial for regulators to perform their enforcement functions. This tiering would also allow the regulatory agency to vet independent researchers for access to platform data that will allow independent audits using information not available to the general public.[90]

A cross-cutting recommendation concerns the need for standards in the reporting of information to the public and in releasing data for the research community. Researchers have been frustrated by the differences in the reporting practices of the different digital platforms, which impede making cross-platform comparisons based on published aggregate data. In the same way, independent research comparing platforms is more informative when the underlying data is released in a standardized, machine-readable format that facilitates comparison. Because the platforms differ in the way they collect, structure and present information to their users, this need for cross-cutting standardization faces enormous practical challenges. A regulatory agency supervising the digital social media platforms could help to coordinate the needed standards-development project. In the absence of a regulatory

program, voluntary efforts among researchers can begin and facilitate the coordination process. Efforts must be made to ensure that the standards facilitate genuine and informative comparisons that take into account the different platform rules and policies.

Scope of Transparency Requirements

To whom do the new transparency requirements apply? To technology companies? Platforms? Social media companies? And within this group, do the requirements apply to all companies, or just to the largest ones? Are the transparency requirements tiered, that is, do they provide for strong measures that apply to large companies and less onerous ones that apply to small- and medium-sized companies?

While these are complex and controversial questions, reasonable decisions are possible in connection with each of them. This paper adopts the position that the **transparency requirements apply to social media companies**. The paradigm cases of these companies are Facebook, YouTube, Twitter, Reddit, 8chan, and so on. The transparency requirements should also apply to search engines because they have their own moderation practices. Any legislative definition needs to capture these cases. A tentative definition of this group of companies might be drawn from existing or proposed laws. For instance, the transparency requirements could apply to "companies that allow users to share or discover user-generated content or interact with each other online," which is the definition used in the UK online harms paper. This definition would include "social media platforms, file hosting sites, public discussion forums, messaging services and search engines."[91] An alternative, based on Senator Mark Warner's (D-VA) proposed pro-competition legislation, might define the scope of transparency requirements to include "consumer-facing communications and information service providers" and include "online messaging, multimedia sharing and social networking."[92] A third alternative, drawn from the German NetzDG law, might be to apply transparency requirements to companies that "operate internet platforms which are designed to enable users to share any content with other users or to make such content available to the public (social networks)."[93] The precise terms of the definition would need to be further developed in the course of developing and processing specific legislative proposals.[94]

Private social media services such as a company's chat function or shared workspace should not be included in the legislative definition. Inevitably, there will be borderline cases, and legislation should provide the regulatory agency with sufficient, but constrained, discretion to adjudicate them.

The paper also adopts the view that the transparency requirements **should apply only to the largest social media companies**. These companies are the ones that are subject to widespread and increasing public concern in connection with their content moderation practices. They are where the largest audiences are to be found and where the failure to provide good content moderation and the correlative failure to adequately protect freedom of expression will create the greatest harm. A reasonable cut-off will have to be based on the number of users within a jurisdiction and will need to be relative to the size and scale of the market in that jurisdiction. For instance, the requirements of NetzDG do not apply to a social network that "has fewer than two million registered users in the Federal Republic of Germany." Senator Warner's proposed law applies its strongest requirements to large communications providers that have "more than 100,000,000 monthly active users in the United States."[95] Each jurisdiction will have to make that determination for itself.

Still, the spread of harmful material and the harms caused by secret moderation techniques already take place on smaller platforms and are likely to increase as the largest platforms improve their

disclosure practices under the pressure of regulatory supervision. These displacement effects of large company regulation are real, as problematic users move from large platforms to smaller ones to avoid platform disclosure rules. The paper proposes to deal with this likely development through the recommendation that the regulatory agency created to implement and supervise transparency rules also have the residual authority to **extend these requirements to smaller and medium-size companies as needed** to achieve the objectives set out in the transparency law. The agency would also have the authority to impose various tiered obligations on companies of different sizes. Rather than permanently limiting the agency to implementing a uniform set of transparency rules just for large companies, the enabling legislation should provide substantial residual authority to expand and tier regulations as the marketplace evolves.

Disclosures to Users Concerning the Operation of Content Moderation Programs

i. Platform rules

All major platforms already provide public disclosures of their content rules and in some cases the enforcement guidance interpreting the standards. Platforms should go further and release their enforcement guidelines along with the policies. This would provide needed insight into the reasons rules are made and enforced, similar to legislative history for new laws or written opinions by courts. Rules and policies must be supported by transparency to be legitimate in the eyes of broader society. Similarly, changes to platform content rules and enforcement guidelines should be communicated to users in a clear, conspicuous, and timely fashion.

Some platforms such as Reddit provide a history of changes in their privacy rules through a system of dropdown tabs that identifies their evolution over time.[96] Wikipedia, though not a social media platform as the term is used in this paper, provides a history of the evolution of its terms of service.[97] While this paper does not recommend this system as a requirement for all platforms, platforms might consider adopting such historical disclosures voluntarily. It is the kind of requirement that might prove to be valuable over time and should be on the list of policy tools available to the supervising regulator.

ii. Range of Enforcement Techniques

Platforms have a range of enforcement techniques to deal with violations of their content rules. Content can be delayed, demoted, or deleted depending on the nature, severity, or frequency of the offense. Accounts can be suspended temporarily or permanently.

Platforms have internal standards for making the judgment about which enforcement action is needed in particular cases. They should provide users and the public with appropriate access to these internal standards. This would prevent arbitrary and capricious treatment of certain users and help to expose different standards of enforcement that might be imposed on different groups of users.

iii. Complaint Procedures

On all the major digital platforms, users can flag a post through relatively easy-to-use options that appear alongside the post itself. By selecting one of these options, a user can report the post as a

violation of the platform standards and identify the type of violation. The complaint and associated post are routed through an automated system that determines how it should be reviewed. If this automated system determines that the content is clearly a violation, then it may be automatically removed. If the system is uncertain about whether the content is a violation, the content is routed to a human reviewer.

This process should be more clearly explained to users who file a complaint, as well as any follow-up procedures that complainants may use if their complaint is rejected or the enforcement action is not appropriate in their judgment.

iv. How platforms explain their decisions

Platforms currently send users whose content has been deleted, delayed, or demoted a message saying that the content violates a community rule. To improve transparency, platforms should cite the specific provision of their rules that the post violated, and why the content was thought to violate that provision. They should provide a link to that provision and to the enforcement guidelines related to that specific provision.

Platforms sometimes respond to users who complain that a post violates a community standard. Platforms should respond to all complaints, letting complainants know what has been done in connection with the complaint. If the post is left up because the platform has judged that it does not violate community standards, the platform should provide the complaining user with an explanation of why it was found permissible.

If the post has been demoted or its distribution delayed or restricted, the platforms should explain both to the complaining party and to the user whose content was affected why that enforcement action was selected rather than any other. Platforms should take necessary steps to protect the complaining party from retaliation or other abuse resulting from the complaint. If there are opportunities to ask for further review of the material, these opportunities should be clearly explained at the time the platform responds to the user's complaint.

v. Appeals process

Platforms currently notify users when their content is no longer available to other viewers and offer users an opportunity to request a review. Platforms should provide users who request a review the opportunity to explain why their content did not violate the community standard cited. In its response to the user, the platform should acknowledge these points and reply appropriately.

Most platforms do not provide complaining users with the opportunity to seek further review in cases where the initial complaint is denied, or the enforcement action is thought to be inadequate. Platforms should provide this additional opportunity for review and inform users of these opportunities at the time they respond to the initial complaint.

<u>Enhanced Public Reporting</u>

Platforms release regular transparency reports in part to respond to legal requirements, such as the reporting obligation in NetzDG, and in part to respond to public concern about the extent, rationale, and effectiveness of their content moderation programs. Platforms should maintain these disclosure programs and improve them along the following lines.

i. Accuracy of content moderation enforcement efforts

Platforms typically screen all content via matching algorithms to reidentify content which has already been identified as inappropriate and fingerprinted in a database for that purpose. For instance, platforms consult their own fingerprinted databases of content that previously was found to be terrorist material or child exploitation, and check hashtag databases maintained by external organizations such as GIFCT for terrorist material[98] and the Internet Watch Foundation for child exploitation images.[99] After this initial screening, the material is posted and subsequently subjected to further automated screening using systems deemed sufficiently reliable to determine likely violation of standards. If the automated system shows a clear violation, such as material that is highly likely to be nudity or a new child exploitation image or fresh terrorist content, the material is removed without further human review. If the judgment is uncertain, or if the material has been flagged by a user as a potential violation, then it is routed to a human reviewer.[100] Platforms sometimes sample reviewer decisions and subject them to further review to determine the "correct" decision.

Some platforms allow users to ask for a second review of material that has been removed, which can result in restoring the material to the site. Platforms generally do not offer a second review when a user flags a post as violating, but the platform decides to leave the post up.

Some platforms publish standard measures of accuracy of their automated detection and removal systems. For instance, they publish "recall," the percentage of posts that were correctly labeled by automated systems as violations out of all the posts that are actually violations. But these accuracy measures are not broken down by type of violation. Because of this, readers have no way of knowing the accuracy rates of automated systems for different types of content (whether the systems are very good at detecting pornography, for example, but struggle with hate speech). Further, the platforms do not publish measures of the accuracy of their human reviewers, which they could do through a reassessment of a sample of human reviews. They typically do not publish reversal rates, although Facebook did do so for the first time in its 2019 report on enforcing its community standards.[101]

Platforms should publish accuracy rates for human reviews, break down the standard accuracy measures to reveal the true and false positives measures on which they are based, and disclose the reversal rates based on a second human review. In addition, these measurements should be available by type of violation, keyed to infringement of specific platform content rules. For human-based content moderation, this level of disclosure will mean that the platforms will have to develop and formalize internal processes and policies to meet a standard of auditability by an external independent auditor. These improvements will give the public a clearer picture of the effectiveness of the enforcement programs.

ii. Reporting the extent of content violations on platforms

Major platforms already publish statistics on the content that violates their community standards or local law. These statistics are made public through voluntary reports, reports issued in conjunction with industry-government collaborations on codes of practice, and mandated transparency reports.

Some platforms report the prevalence of content violating community standards or local law as a percentage of all content viewed. They should consider an additional prevalence metric: the number of "bad" posts in comparison to the number of total posts. It is important to include both as a percentage of viewed posts and as a percentage of all posts. The number of bad posts viewed is affected by recommendation and prioritization algorithms. It is also affected by the effectiveness of automated removal systems that proactively detect violating content and block it before it is posted.

It would help the public to understand the extent to which the input into the platform is problematic rather than just measure the output to the viewers. This would provide an assessment, at a point in time and over time, of the propensity of platform users to violate specific content rules and allow correlations to outside "triggering events" in the real world such as a terrorist incident. It would also provide a way to assess the role of content ordering and moderation systems in blocking or disseminating violating content within the platform.

iii. Reporting on actions taken in response to complaints

Digital platforms often report the actions they have taken as the number of posts or accounts for which they have taken any content moderation step at all, such as blocking a photo or downgrading the material in recommendation engines or removing an account.

Platforms should report the content moderation actions they take broken down by the type of action taken. This would provide an understanding of their propensity to use a severe enforcement action such as account deletion in contrast to a milder action such as downgrading the content. This would be especially valuable if provided by type of violation such as hate speech versus terrorist material. In addition, the number of actions should be reported as a percentage of all posts or accounts involving violating content. This provides a picture of the effectiveness of the enforcement effort and the relative importance of different enforcement techniques. In addition, platforms should report the actions taken by the number of users or accounts involved, discounting the fake accounts. This would provide a sense of whether the source of the violating content is a large percentage of users or accounts or whether a small fraction of users or accounts create the bulk of the content moderation issue.

iv. Measures of effectiveness of content moderation programs

Platforms often report how much violating content is detected and what action is taken before users report it. Facebook calls this reporting an assessment of its proactivity, and it is measured as the platform-detected violating content as a percentage of all content the platform ultimately determines has violated a standard, which includes both the automatically detected content and the content reported by users. So, for instance, of the nudity that was ultimately removed, the platform might have detected and removed 95% before users saw it, leaving only 5% that was reported by users and removed by the platform.

But this metric is potentially misleading. A high percentage in this proactivity measure might lead readers to conclude that the automated systems are very effective. But the metric does not record the

content missed by both the automated system and the users. To get a better estimate of effectiveness, a platform should first estimate the total amount of content that violate its rules, which it can do through a sampling and review procedure independent of its enforcement process. The amount of content that the platform detects before users do can be presented as a percentage of this estimate of all content violating standards.

v. Additional information to achieve outward transparency

In addition to the measures listed above, the mandate for issuing public reports should focus on outward transparency by requiring social networks to disclose essential information on how they operate their core functions. This should include (i) how they rank, organize and present user-generated content; (ii) how they target users with unsolicited content, at their own initiative or on behalf of third parties, usually as a paid service; and (iii) how they moderate the content being published on their platform. The regulator will prescribe the details of what level and structure of information is required to ensure that the relevant information needed for outside audits is presented to the public.

Outward transparency should rely on the obligation to release and maintain up-to-date reference documents on each core function including ranking, targeting, and moderation. These documents should be released in a timely manner, without undue delay, so that researchers and regulators can make use of them while they are still relevant. Platforms should disclose the core structure of the algorithms and how they were developed or trained for machine learning algorithms. The information disclosed should be sufficient to allow an expert to advise policy makers and civil society representatives who engage in the open policy dialogue on substantive issues.

In some instances, disclosing some feature of the platform algorithms could create opportunities for malicious third parties to circumvent or abuse platform security and potentially harm or mislead its users. In those circumstances, regulators should work with platforms to ensure that whatever data is released is both sufficiently transparent and adequately protects the security of the platform. A robust and transparent waiver program would enable such a dialogue to take place.

The regulator should have the authority to prescribe and adjust over time the depth and scope of the disclosure requirements, based on the needs arising from the open policy dialogue and following public consultation of all interested parties. In addition, the regulator should have investigative powers to cross-check the accuracy of the information released to assure the public that the information accurately reflects the underlying processes.

Access to Data for Researchers and Regulators

Disclosures to users and public reports can provide essential elements of transparency, which provides value in its own right as an accountability measure and a means to enable additional accountability measures. But the value of disclosures relies on a level of trust in the platforms themselves that is lacking in the current climate of opinion, even when regulators or researchers have tools to check the validity of information released. Good governance, moreover, should rely as little as possible on trust. Public companies, for example, rely on external auditors who have access to their books and records to reassure investors concerning their financial health. In a similar way, digital platforms must open

their operations to an appropriate degree in order to assure the public that their systems are functioning properly.

In addition, algorithms, especially machine learning algorithms, may exacerbate unintended biases that are not known by the company itself and thus are not captured and disclosed under an outward transparency scheme. These biases can often only be revealed by an active scrutiny review.

For these reasons, as an additional transparency measure, researchers and regulators should have access to platform data to audit the systems involved and assure the public that they are operating as intended and without unintended bias. These disclosed assessments would enable a public judgment concerning whether the companies are operating in the public interest in connection with their content moderation activities. This section describes this paper's recommendations for access to data.

This type of inward transparency should rely on the obligation of the major social networks to develop, at their own cost, a secured platform for accredited outside researchers to access the needed data to implement research of general interest, implement the needed data processing, and extract the results without compromising the private data of users and the value of the aggregate data of the social network.

The process should be supervised by an independent regulator in charge of:

- Defining priorities for research of general interest, following a public consultation and based on the policy dialogue of substantive issues arising from social network operations.
- Organizing the process through which academics can apply for access to the platform. The platform itself should not decide on the merits of the research considered but rely on peer review committees following academic standards and set up under the supervision of the regulator. The social networks should have the opportunity to comment on the proposed research project.
- Settling disputes between social network and academics that arise from the implementation of this controlled access.

A basic presumption of these recommendations is that a system of tiered access is essential. Some data needs to be widely available to the general public and freely available to researchers to conduct whatever research they deem important and worthwhile. Other data might be sensitive, for content, privacy or commercial reasons, and this material should be restricted to researchers vetted by or working with the supervising regulator.

Businesses spend resources to collect and organize data concerning their own systems for their own business purposes. The requirement for transparency in connection with these systems of business records should not, in general, impose an obligation on the platforms to develop or collect new data. The data they need should already be available within their management systems. The needs of transparency might require that the data be organized or sorted or presented to the outside world in ways that go beyond business needs. To some degree platforms are doing this already when they construct their transparency reports. The recommendations in this section are designed to provide a reasonable level of transparency for outside researchers and regulators without an undue burden on the platforms themselves.

i. Access to data on the operation of content moderation programs

In addition to the metrics that platforms themselves publish, platforms need to improve the amount, nature, and format of information on the operation of their content moderation programs they provide to outside researchers and regulators so as to allow comprehensive audits.

A key element of the successful operation of content moderation programs is an effective and efficient complaint process. This paper recommends that platforms commit to archiving complaints, allowing third-party oversight of issues that have triggered user complaints and the record of the platforms in responding to these complaints. This disclosure information should include the complaint itself, the content that was the subject of the complaint, the action that was taken or not taken in response to the complaint, the alleged rule violation, the time it took to respond to the complaint, whether a second review was requested, and the outcome of any second review. The material would need to be in machine-readable format to facilitate computational analysis. It should be searchable by anonymized ID, date, nature of content and content rule (allegedly) violated. To protect the privacy of users, all complaint data should be anonymized using reasonable techniques such as k-anonymity or differential privacy. All users of the archive should be under a contractual obligation to avoid all attempts to reidentify the individuals involved and should be subject to suspension of their access to the complaint archive for violation of the prohibition on reidentification.

In addition, researchers need the underlying data upon which published estimates of errors are based. This applies to the algorithms that are used for initial screening, as well as the algorithms that are used to identify content that is likely to violate platform rules. It also applies to initial human reviews, further reviews as requested by complaining users or users whose content has been deleted or downgraded, and the reviews of samples of moderated content that are used to establish an internal baseline of accuracy.

The platforms should develop a mechanism to make disaggregated data on the prevalence of violating content available to third-party researchers. In this way, outside researchers will be able to duplicate the published platform aggregate data, thereby increasing trust in the reported results. Platforms should preserve consumer privacy in releasing this information using protective statistical techniques such as k-anonymity or differential privacy. In meeting this challenge of balancing transparency and individual privacy, platforms can be guided by the experience of statistical agencies such as the U.S. Census Bureau.

In some cases, such as those involving terrorist material or child exploitation images, public disclosure of the content taken down would be counterproductive. In these and in other cases where auditing is important but public disclosure problematic, platforms should retain copies of the relevant information for review by a supervising government agency, with access provided only to researchers approved by the agency. In cases where privacy interests or commercial secrets are of the utmost concern, reasonable privacy safeguards backed by contractual obligations might not be sufficient to protect these interests. In these cases as well, information can be supplied to the agency and made available only to approved researchers.

ii. Political advertising

Political advertising can be a major vector for disinformation campaigns, which have the potential to disrupt and challenge the integrity of democratic processes. In response to this challenge, platforms have committed to creating archives of political ads for researchers to access in real time. The hope is

to detect advertising campaigns that aim to disrupt elections in a timely fashion and to conduct long-term research to reveal the methods used so as to guard against them more effectively in the future.

Despite much progress in this area and good intentions on the part of the platforms, researchers and other commentators have found significant flaws in these archives of political ads.[102] They recommend significant changes to improve the flow of information and to encourage, not limit, research.

In connection with voluntary efforts to provide access to platform data about political ads, this paper generally endorses the recommendations of the Mozilla Foundation.[103] The types of ads covered should include electioneering content, ads concerning candidates or holders of political office, matters of legislation or decisions of a court, and functions of government. The information disclosed should cover the content of the ad, the targeting criteria, the number of impressions, user engagement beyond viewing the ad, and the price paid to place the ad. The method of disclosure should provide unique identifiers for the ads and advertiser, machine-readable access, the ability to quickly download large amounts of data in a timely fashion, including historical data, and search capability by ad content, author and date. Platforms should make political ads available within 24 hours of publication, maintain access going back 10 years, and create programming interfaces to allow long-term studies.

As with all efforts to improve disclosures, the details of these voluntary efforts need to be worked out cooperatively with the platforms and the research community. A priority should be the establishment of an institutional outreach structure that allows modification of access functionality as the nature of political ads changes and research needs evolve.

Nevertheless, the nature of these disclosures should not be limited to what the platforms can work out with researchers. Policy makers should establish a floor for adequate disclosure to ensure that a minimum of needed information is available to conduct adequate audits of the use of platforms for political advertising purposes.

This paper recommends that legislatures require public disclosures in connection with political ads on platforms. In particular, it is generally supportive of the requirements of the Honest Ads Act, believing that the disclosures required by that legislation would be a meaningful start to better platform transparency.[104]

Platforms must disclose information about advertisements urging the election or defeat of candidates for public office and paid political issue-oriented ads in a publicly accessible database in a machine-readable format.

The ads to be included in this database are reasonably defined in the Honest Ads Act, focusing on any advertisement made by a candidate or that communicates a message relating to "any political matter of national importance" which includes "a candidate," "any election to Federal office," or "a national legislative issue of public importance." The file should be maintained for a period of years sufficient to allow retrospective research. Many of the details of the terms of access and search capabilities are complex technical issues that would need to be sorted out in a public rulemaking by an expert agency, but should provide at a minimum that researchers be able to search the database by candidate name, issue, purchaser and date.

This paper recommends covering issue ads when the sponsor pays the platform for enhanced distribution or targeting. They influence the political conversation and can directly or indirectly affect the outcome of elections. But advocacy activity on issues of public importance that do not involve payment to the platforms would not be covered by requirements for disclosure of political ads. If

advocacy activity not involving payment to social media platforms needs to be regulated to ensure authenticity, this must be done separately from requirements for disclosure of political advertising.

The information needed in the file would include: a copy of the advertisement, a description of the audience targeted, the number of views generated from the advertisement, and the date and time that the advertisement is first displayed and last displayed; the average rate charged for the advertisement; the name of the candidate, the office sought or the national legislative issue involved; and information about the purchaser of the ad. A crucial element is that both targeting information and audience information needs to be disclosed.

Finally, the agency involved in supervising the mandated disclosure requirements should have ongoing regulatory responsibility for the conduct of platforms in connection with political ads in much the same way that the Federal Communications Commission in the United States maintained its supervisory role over the required broadcasting and cable disclosures concerning political ads. In conjunction with this supervisory role, the agency should have broad powers to access information for enforcement purposes that might not be made available to the general public or to scholarly researchers.

This agency collection and use of platform information for enforcement purposes should be carefully crafted to prevent agency coercion of platforms or political actors for the political ends of the agency itself or the political party that happens to be in charge of the government. The agency should be prohibited from reaching into the activities of the platforms to direct or dictate a political outcome or to gather intelligence to be used to favor some political actors over others. The agency would need to conduct public rulemakings with court review to ensure consistency with the authorizing statute and to prevent arbitrary and capricious action. The rulemakings should also determine the types of information to be collected for disclosure enforcement purposes, the measures to ensure that platform information warranting confidentiality is not revealed to the public, and the oversight mechanisms to protect against political abuse of the agency's enforcement powers. Some possible agency uses of platform information for enforcement activities are listed below.

Regulators might seek to conduct their own research through in-house experts or specialist contractors to verify the real identities of political advertisers who do not fully disclose who they are when they buy ads, and to put into place identity verification requirements to mitigate the risks of misidentification. Such minimum verification requirements could build on the systems some platforms already have in place and would have the advantage of uniformity and the legitimacy of action based on a democratic mandate from a legislative or regulatory body.[105]

The agency would also need regulatory and research powers to investigate the extent to which native advertising techniques can be used to escape political ad disclosure. When a political advertiser pays a sponsorship fee to a platform for distributing editorial content, this might not be included in the list of political ads disclosed in an ad archive. Platforms and the regulatory agency would need to work together to find a way to identify and include these paid efforts to influence the political landscape in political ad disclosures.

Targeting criteria used by platforms need to be disclosed to the public, but the level of granularity involved can jeopardize user privacy, since advertisers sometimes target their campaigns based on personal information such as email address or telephone number. The trade-off between transparency and user privacy cannot always be specified in advance and might need ongoing supervision by the regulatory agency involved, and consultation with data protection authorities.

iii. Content-ordering techniques

Algorithms determine the priority of content delivered to platform users and construct recommendations for users to explore further content. The basis for these content-ordering techniques is unclear, but they seem designed to maximize attention or user engagement with the platform, without regard to substantive content. As a result, critics have alleged that these algorithmic ordering techniques can lead users into further exploration of terrorist material, disinformation campaigns and material promoted by hate groups.

In principle, these same techniques could also be designed to pursue a political objective. Platforms have the capacity to use content-ordering techniques to promote certain ideas. One can imagine that a platform might one day decide to increase the visibility of certain content, increasing, for instance, the awareness of climate change or promoting their preferred course of action in connection with a public policy issue. As part of the principle of freedom of expression, they should be permitted to do so, subject to any applicable regulations.

Information about these algorithms is needed to audit their role in disseminating and amplifying problematic content, or simply in influencing the public debate and the formation of public opinion. In the latter case, such disclosure is a direct counterpart of the freedom they enjoy and the corresponding accountability principle. Internal reviews are key, but it is important that independent researchers and regulators have sufficient access to these algorithmic techniques to evaluate their role in increasing the distribution and salience of problematic content or platform-preferred political content, and to recommend or establish measures to reduce the prevalence of this material or otherwise to regulate it.

Revelation of the source code or formula of the relevant algorithm is widely viewed as irrelevant. The key auditing measures are disclosure of the aim being pursued in designing the algorithm, input-output analysis to assess unintended effect, and an understanding of the key factors at work in recommendation and personalization algorithms.[106] For this reason, enough information has to be available to outside researchers to enable them to conduct these audits.

Audits based on information available to the public might necessarily be more limited than audits based on all the information available to the platform itself. Platforms have access to the formula used in content moderation and content-ordering algorithms and can use that information to troubleshoot. But an input-output study is still possible as demonstrated by the Upturn study cited earlier. Similar, external testing might be able to detect recommendation and personalization outputs that privilege hate speech, disinformation campaigns or terrorist material, or actively promote a legitimate opinion.

When a platform seeks to adjust an algorithm, it should publish enough information about the change to allow outside researchers to assess the implications of the change. For instance, when YouTube recently adjusted its engagement algorithm it released an assessment of the changes, but not enough information for researchers to understand whether the changes would make the algorithm better at recommending more addictive content or better at controlling rabbit holes of hate speech and terrorist material.[107]

In addition, the public and the regulator need an understanding of the major factors at work in the operation of these algorithms. This need not be the formula or source code, but rather a description of the key factors driving the operation of the classifiers that govern the content-ordering functions.

In the credit-granting context, the provision of explanations is standard procedure and has been for generations. Credit-granting institutions and the service providers that furnish risk assessment tools have built into their systems and business models the capacity to respond to the regulatory requirements for notices in connection with adverse actions that list the major factors involved in denying a loan or providing it with more stringent terms and conditions.[108]

For this reason, this paper recommends that the regulator have the power to request explanations about the way algorithms operate, and to require platforms to provide these explanations in appropriate form to platform users.

Transparency also requires clarity about the purposes or objectives of algorithm optimization. In other contexts, it is clear what is being predicted – for credit decisions, for instance, the lender wants to know the probability of default. But the objective of content-ordering algorithms is not clear to the user, to the regulator or to the outside auditing researcher.

The input data is a crucial element needed for successful audits. As ISD has recommended, "The regulator should be able to identify and assess what data was used to train the algorithm, how it was collected, and whether it is enriched with other data sources, and whether that data changed over time."[109]

Conclusion

The recommendations in this paper are designed to further the public's interest in the transparent operation of digital social media platforms. This transparency aims to ensure that these platforms provide both effective content moderation and a robust environment for free expression on crucial matters of public importance. Some of the recommendations are detailed and focused on technical measures for assessing and presenting clearly how content moderation systems work. But the general thrust is more important than any of the detailed recommendations.

The transparency measures described in this report find their natural and most effective home in a supervising regulatory agency with authority to enforce, implement and upgrade this regulatory structure, including its transparency requirements.

Still, better should not be the enemy of good. Much can be done without legislation and the oversight of a dedicated supervisory agency. This paper's call for further legislation should not be interpreted as a recommendation to end or curtail the valuable voluntary efforts that the major platforms are pursuing. On the contrary, many of its specific recommendations can be incorporated into these ongoing efforts.

The key area for disclosure is the content moderation system itself, especially concerning how users can take advantage of a platform's complaint process. This will be of vital concern to many users who want to preserve a digital social media platform free of harmful material, but one that allows them full freedom to express views on controversial issues. A second level of disclosure consists of public reports describing for both users and subject matter experts how the relevant internal systems are performing. These reports need to contain enough detail for experts to understand the systems and make recommendations to governments on improvements. Finally, there need to be data disclosures to researchers and regulators to enable audits. This last level needs tiering to protect other interests,

including privacy and commercial secrets, so that access is provided only to vetted researchers when needed.

The recommendations for transparency are intended to set out a gradual, pragmatic and proportionate approach for governance of digital social media platforms. They will allow the regulator to build and adjust the enhanced transparency standard over time with constant feedback loops. This allows the regulator to adapt the scope and the depth of the disclosure to the evolution of substantive issues, based on open policy dialogue. The approach requires the regulator to provide guidance and eventually to decide upon the needed limits to such transparency requirement and the trade-off between the general interest for such disclosure, the privacy interests of platform users, and the private interest of the social networks, including, inter alia, the adversarial impact of some disclosure, which could jeopardize the integrity of social networks.

The transparency system outlined here is not the only regulatory measure that might need to be taken. But it is a crucial first step that will provide needed information for open and informed policy discussions. Moreover, it has a key advantage over more intrusive content-based measures. It limits the dangers to democratic self-governance that arise when government agencies are able to control the flow of information citizens rely upon for making democratic decisions.

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. Thanks to Jeff Gary, Associate at the Institute for Technology Law and Policy at Georgetown Law, for invaluable research assistance. Thanks also to the members of the Transatlantic Working Group on Content Moderation and Freedom of Expression for their helpful comments in connection with the Bellagio, Italy Session, November 13-16, 2019. This paper does not necessarily reflect the views of the group or any individual member of it.

[2] Thomas Wischmeyer, 'What is illegal offline is also illegal online' – The German Network Enforcement Act 2017 in B. Petkova and T. Ojanen (eds.), *Fundamental Rights Protection Online: the Future Regulation of Intermediaries*, Edward Elgar, forthcoming 2019. Heidi Tworek and Paddy Leerssen, An Analysis of Germany's NetzDG Law, Working Paper, Transatlantic Working Group on Content Moderation Online and Freedom of Expression April 15, 2019) https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf. The English text of NetzDG is available here: https://germanlawarchive.iuscomp.org/?p=1245. It is now considering new legislation establishing various non-discrimination and transparency obligations for video platforms like Netflix and for media intermediaries like YouTube. See Natali Helberger, Paddy Leerssen and Max Van Drunen, Germany proposes Europe's first diversity rules for social media platforms, LSE Blog, May 29, 2019, https://blogs.lse.ac.uk/medialse/2019/05/29/germany-proposes-europes-first-diversity-rules-for-social-media-platforms/.

[3] "France's lower house passes online hate speech law" (France24, July 9, 2019, https://www.france24.com/en/20190709-frances-lower-house-passes-online-hate-speech-law) obliging social media networks to remove hate speech in 24 hours. Sites that fail to comply with the law by not removing "obviously hateful" content risk fines of up to $1.4 million.

[4] European Parliament, "Terrorist content online should be removed within one hour, says EP" (*European Parliament Press Room*, April 17, 2019, http://www.europarl.europa.eu/news/en/press-room/20190410IPR37571/terrorist-content-online-should-be-removed-within-one-hour-says-ep); Joris van Hoboken, The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications, Working Paper,

Transatlantic Working Group on Content Moderation and Freedom of Expression, May 3, 2019, https://www.ivir.nl/publicaties/download/TERREG_FoE-ANALYSIS.pdf.

[5] United Kingdom, Online Harms White Paper (Department for Digital, Media, Culture & Sport, April 30, 2019) paragraph 16, available at https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper-executive-summary--2#contents. This proposal would impose a duty of care on platforms obligating them to remove harmful content or face substantial fines. See Peter Pomerantsev, A Cycle of Censorship: The UK White Paper on Online Harms and the Dangers of Regulating Disinformation, October 1, 2019, https://www.ivir.nl/publicaties/download/Cycle_Censorship_Pomerantsev_Oct_2019.pdf.

[6] Parliament of Australia, Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019, April 5, 2019, https://www.aph.gov.au/Parliamentary_Business/Bills_Legislation/Bills_Search_Results/Result?bId=s1201.

[7] Republic of Singapore, Government Gazette Acts Supplement, Protection from Online Falsehoods and Manipulation Act 2019, June 28, 2019, https://sso.agc.gov.sg/Acts-Supp/18-2019/Published/20190625?DocDate=20190625.

[8] The approach in this paper owes much to the recommendations in 'Creating a French framework to make social media platforms more accountable: Acting in France with a European vision' (Final Mission Report on the Regulation of social networks – Facebook experiment, submitted to the French Secretary of State for Digital Affairs, May 2019), https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf.

[9] This paper is agnostic whether the supervisory agency is a brand new entity or whether it is part of an existing regulatory agency such as, in the United States, the Federal Communications Commission or the Federal Trade Commission or in Europe media regulation agencies such as CSA in France or OFCOM in the United Kingdom.

[10] Arunesh Mathur, et al., Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites, September 20, 2010, https://arxiv.org/pdf/1907.07032.pdf.

[11] The U.S. Federal Trade Commission has experience in treating commitments by industry as enforceable promises and has taken action against companies who violate these pledges in cases concerning the privacy shield commitments by U.S. companies to abide by certain privacy practices and also in cases involving company commitments to certain privacy practices in the area of the privacy of student information.

[12] See the description of FINRA's structure and mode of operation at their website, https://www.finra.org/#/.

[13] Jacinda Ardern, Christchurch Call to eliminate terrorist and violent extremist online content adopted, Press Release, May 16, 2019, https://www.beehive.govt.nz/release/christchurch-call-eliminate-terrorist-and-violent-extremist-online-content-adopted.

[14] Electronic Frontier Foundation and others, "Background Paper on the Manila Principles on Intermediary Liability," ManilaPrinciples.org, May 30, 2015, https://www.eff.org/files/2015/07/08/manila_principles_background_paper.pdf#page=49.

[15] The Santa Clara Principles on Transparency and Accountability in Content Moderation, May 7, 2018, https://santaclaraprinciples.org.

[16] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/HRC/38/35 (United Nations Human Rights Council, April 6, 2018), https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/096/72/PDF/G1809672.pdf.

[17] Danielle Keats Citron and Frank A. Pasquale, "The Scored Society: Due Process for Automated Predictions" (2014) 89 Washington Law Review 1, http://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1318/89WLR0001.pdf?sequence=1.

[18] Brent Harris, "Establishing Structure and Governance for an Independent Oversight Board," Facebook, September 17, 2019, https://newsroom.fb.com/news/2019/09/oversight-board-structure/.

[19] Jeff Jarvis, "Proposals for Reasonable Technology Regulation and an Internet Court," Medium, April 1, 2019, https://medium.com/whither-news/proposals-for-reasonable-technology-regulation-and-an-internet-court-58ac99bec420.

[20] "Social Media Councils: From Concept to Reality," Stanford Global Digital Policy Incubator, ARTICLE 19, and David Kaye, UN Special Rapporteur on the Right to Freedom of Opinion and Expression (February 2019), https://cyber.fsi.stanford.edu/gdpi/content/social-media-councils-concept-reality-conference-report; Heidi Tworek, Social Media Councils, CIGI, October 28, 2019, https://www.cigionline.org/articles/social-media-councils.

[21] Facebook is beginning to provide these reporter appeals. In its November 2019 Community Guidelines Enforcement Report (https://transparency.facebook.com/community-standards-enforcement/guide), Facebook says, "We are beginning to provide appeals not just for content that we took action on, but also for content that was reported but not acted on. These reporter appeals are not included in the report."

[22] These laws are enforced in the United States by the Securities and Exchange Commission. See the list of U.S. security laws at https://www.sec.gov/answers/about-lawsshtml.html.

[23] The U.S. Federal Trade Commission enforces Section 5 of the Federal Trade Commission Act, 15 U.S.C. § 45(b).

[24] For instance, disclosure that Cloudflare provided service to white nationalists and other prompted them to rethink their provision of service to 8Chan, a social media service that was instrumental in spreading the Christ Church video and provided inspiration for the shooter who killed 20 people in an El Paso Wal-Mart. See Matthew Prince, "Terminating Service for 8Chan," Cloudflare Blog, August 4, 2019, https://blog.cloudflare.com/terminating-service-for-8chan/.

[25] Heidi Tworek, "Social Media Platforms and the Upside of Ignorance," CIGI, September 9, 2019, https://www.cigionline.org/articles/social-media-platforms-and-upside-ignorance.

[26] An outside group discovered a disparate impact in Facebook's delivery of housing ads, which potentially violates existing anti-discrimination rules. See Adi Robertson, "Facebook's ad delivery could be inherently discriminatory, researchers say," The Verge, April 4, 2019, https://www.theverge.com/2019/4/4/18295190/facebook-ad-delivery-housing-job-race-gender-bias-study-northeastern-upturn. The Upturn study can be found here: https://arxiv.org/pdf/1904.02095.pdf.

[27] Senator Josh Hawley (R-MO) introduced legislation to make Section 230 liability protections conditional on certification of political neutrality in content moderation by the Federal Trade Commission.

[28] It should be a source of concern to those urging content-based restrictions that countries whose traditions do not emphasize liberal values and individual rights are at the forefront of these measures. As noted before Singapore has a new strong law against propagating "false statements of fact." Newspapers in China are urging Hong Kong to adopt a similar measure to deal with the "fake news" on traditional media and social networks concerning the Hong Kong demonstrations. See "Hong Kong Needs to Fight Fake News through Legislation," Global Times, October 8, 2019, http://www.globaltimes.cn/content/1166303.shtml. Under legislation proposed in Russia in October 2019, platform companies would have to take down illegal content and block users posting it within 24 hours if asked to do so by the state communications agency. See Reuters, "Russian Lawmakers Look to Ban Email Users Who Share Illegal Content," The Moscow Times, October 9, 2019, https://www.themoscowtimes.com/2019/10/09/russian-lawmakers-look-to-ban-e-mail-users-who-share-illegal-content-a67657.

[29] Adam Satariano, "Facebook Can Be Forced to Delete Content Worldwide, E.U.'s Top Court Rules: The decision that individual countries can order Facebook to take down posts globally sets a benchmark for the reach of European laws governing the internet," New York Times, October 3, 2019, https://www.nytimes.com/2019/10/03/technology/facebook-europe.html. See also Court of Justice of the European Union, Judgment of the Court, C-18/18, *Glawischnig-Piesczek v. Facebook*, http://curia.europa.eu/juris/document/document.jsf?text=&docid=214686&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=4239414. Facebook observes that this ruling "undermines the long-standing principle that one country does not have the right to impose its laws on another country." Monika Bickert, European Court Ruling Raises Questions about Policing Speech, Facebook, October 14, 2019, https://about.fb.com/news/2019/10/european-court-ruling-raises-questions-about-policing-speech/.

[30] Sarah Marsh, "'Right to be forgotten' on Google only applies in EU, court rules: Europe's top court says firm does not have to take sensitive information off global search," The Guardian, September 24, 2019, https://www.theguardian.com/technology/2019/sep/24/victory-for-google-in-landmark-right-to-be-forgotten-case. See also Court of Justice of the European Union, Judgment of the Court, C-507/17, Google v. CNIL, September 24, 2019, http://curia.europa.eu/juris/document/document.jsf;jsessionid=0A97493A2186F3D627007F364D681E11?text=&docid=218105&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=259171.

[31] These issues arose in discussions with members of the Transatlantic Group in Bellagio.

[32] Evan A. Feigenbaum, "In Asia, Disruptive Technonationalism Returns," Carnegie Endowment for International Peace, November 13, 2019, https://carnegieendowment.org/2019/11/13/in-asia-disruptive-technonationalism-returns-pub-80331.

[33] Brandon Pho, "New State Law Requires More Transparency from Social Media Political Ads," Voice of OC, October 3, 2018, https://voiceofoc.org/2018/10/new-state-law-requires-more-transparency-from-social-media-political-ads/. The text of the law can be found here: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB2188.

[34] The law requires those using bots to disclose that the bot is a bot to every person the bot communicates with. The text of the law can be found here: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001.

[35] CCPA has transparency/disclosure requirements focused on data collection and use. See this chart prepared by the International Association of Privacy Professionals: https://iapp.org/resources/article/cacpa-what-to-disclose-and-where-to-disclose-it/.

[36] The disclosure requirements are contained in Article 13 of the GDPR, the text of which can be found here: https://gdpr-info.eu/art-13-gdpr/.

[37] The reporting requirement is in Section 2(1) of NetzDG. See https://germanlawarchive.iuscomp.org/?p=1245.

[38]The full list of these reporting requirements is in Section 2(2) of NetzDG, available at https://germanlawarchive.iuscomp.org/?p=1245.

[39] The networks have complied with this requirement in various ways. See Facebook, https://fbnewsroomus.files.wordpress.com/2018/07/facebook_netzdg_july_2018_english-1.pdf; Instagram, https://instagram-press.com/wp-content/uploads/2019/07/instagram_netzdg_July_2019_english.pdf; Twitter, https://transparency.twitter.com/en/countries/de.html; Google, https://transparencyreport.google.com/netzdg/overview?hl=en_GB.

[40] Section 4(4) of NetzDG. See https://germanlawarchive.iuscomp.org/?p=1245.

[41] Thomas Escritt, "Germany fines Facebook for under-reporting complaints," Reuters, July 2, 2019, https://www.reuters.com/article/us-facebook-germany-fine/germany-fines-facebook-for-under-reporting-complaints-idUSKCN1TX1IC. The press release from the Federal Office of Justice is here: https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html.

[42] Joe Fuld, "What Do New State Laws On Political Digital Ads Mean For You?," Campaign Workshop Blog, November 1, 2018, https://www.thecampaignworkshop.com/political-digital-ads-laws.

[43] Dale Eisman, "Political Advertisers Still Breaking Online Disclosure Rules," Common Cause, February 13, 2018, https://www.commoncause.org/democracy-wire/political-advertisers-still-breaking-disclosure-rules/.

[44] Facebook's community standards are available here: https://www.facebook.com/communitystandards/; Google's community guidelines are here: https://about.google/intl/en_us/community-guidelines/; Twitter's rules and policies are here: https://help.twitter.com/en/rules-and-policies/twitter-rules; Reddit discloses its content policy here: https://www.redditinc.com/policies/content-policy-1.

[45] https://www.facebook.com/zuck/posts/10104874769784071; see also, https://www.theguardian.com/technology/2018/apr/24/facebook-releases-content-moderation-guidelines-secret-rules; the interpretative guidelines are now incorporated directly into the community standards, https://www.facebook.com/communitystandards/introduction/.

[46] Facebook's November 2019 community standards enforcement report is here: https://transparency.facebook.com/community-standards-enforcement; Google's Community Guidelines Enforcement Report for YouTube for the period July - September 2019 is here: https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB.

[47] Donie O'Sullivan, "Twitter cracks down on state media after unveiling Chinese campaign against Hong Kong protesters," CNN Business, August 20, 2019, https://www.cnn.com/2019/08/19/tech/china-social-media-hong-kong-twitter/index.html.

[48] Twitter Safety, "Information operations directed at Hong Kong," August 19, 2019, https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html.

[49] Twitter Inc., "Updating our advertising policies on state media," August 19, 2019, https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html.

[50] See Twitter's ad transparency center here: https://ads.twitter.com/transparency. In November 2019, it updated its advertising policy to exclude ads with political content, that is, "content that references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome." See https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html.

[51] Google's latest transparency report on political advertising on Google, including YouTube, is available here: https://transparencyreport.google.com/political-ads/home?hl=en_GB.

[52] Katie Harbath and Sarah Schiff, "Updates to Ads About Social Issues, Elections or Politics in the US," Facebook,

October 16, 2019, https://newsroom.fb.com/news/2019/08/updates-to-ads-about-social-issues-elections-or-politics-in-the-us/.

[53] European Commission, Code of Practice on Disinformation, September 26, 2018, https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation; Peter Chase, "The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem," Working Paper, Transatlantic Working Group on Content Moderation and Free Expression, August 29, 2019, https://www.ivir.nl/publicaties/download/EU_Code_Practice_Disinformation_Aug_2019.pdf.

[54] European Commission, Fourth intermediate results of the EU Code of Practice against disinformation, May 17, 2019, https://ec.europa.eu/digital-single-market/en/news/fourth-intermediate-results-eu-code-practice-against-disinformation.

[55] See the discussion of this initiative at their website: https://gifct.org/about/. In September 2019, GIFCT announced that it was going to evolve from a consortium of companies to an independent organization with its own executive director and staff. See GIFTC, Next Steps for GIFCT, September 23, 2019, https://gifct.org/press/next-steps-gifct/.

[56] GIFTC, GIFTC Transparency Report, July 2019, https://gifct.org/transparency/. For concerns regarding the adequacy of this transparency report, see Brittan Heller, "Combating Terrorist-Related Content Through AI and Information Sharing," Working Paper, Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, April 26, 2019, https://www.ivir.nl/publicaties/download/Hash_sharing_Heller_April_2019.pdf.

[57] "Creating a French framework to make social media platforms more accountable: Acting in France with a European vision" (Final Mission Report on the Regulation of social networks – Facebook experiment, submitted to the French Secretary of State for Digital Affairs, May 2019), https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf.

[58] Heidi Tworek and Paddy Leerssen, "An Analysis of Germany's NetzDG Law," (Transatlantic Working Group on Content Moderation Online and Freedom of Expression, April 15, 2019, https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf). See also Keller, D. & Leerssen, P. (Forthcoming), "Facts and where to find them: Empirical foundations for policymaking affecting platforms and online speech," in N. Persily & J. Tucker (eds.), *Social Media and Democracy: The State of the Field.*

[59] H.R.2592 - Honest Ads Act, Introduced by Rep. Derek Kilmer, May 8, 2019, https://www.congress.gov/bill/116th-congress/house-bill/2592; S.1356 - Honest Ads Act, Introduced by Introduced by Senator Amy Klobuchar, May 7, 2019, https://www.congress.gov/bill/116th-congress/senate-bill/1356.

[60] S.2125 - Bot Disclosure and Accountability Act of 2019, Introduced by Senator Diane Feinstein, July 16, 2019, https://www.congress.gov/bill/116th-congress/senate-bill/2125.

[61] Institute for Strategic Dialogue, Online Harms White Paper Consultation, 2019, https://www.isdglobal.org/isd-publications/extracts-from-isds-submitted-response-to-the-uk-government-online-harms-white-paper/.

[62] United Kingdom, Online Harms White Paper (Department for Digital, Media, Culture & Sport, April 30, 2019), paragraph 23, https://www.gov.uk/government/consultations/online-harms-white-paper.

[63] Report Of The Facebook Data Transparency Advisory Group, Yale Law School, April 2019 https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf.

[64] Facebook, Community Standards Enforcement Report, November 2019, https://transparency.facebook.com/community-standards-enforcement. For instance, the reports say only "We respond differently depending on the severity, and we may take further action against people who repeatedly violate standards."

[65] Trevor Timm, "Prominent Security Researchers, Academics, and Lawyers Demand Congress Reform the CFAA and Support Aaron's Law," Electronic Frontier Foundation, August 2, 2013, https://www.eff.org/deeplinks/2013/08/letter.

[66] National Association of Criminal Defense Lawyers, CFAA Cases, April 25, 2019, https://www.nacdl.org/Content/CFAACases.

[67] Timothy B. Lee, "Web scraping doesn't violate anti-hacking law, appeals court rules: Employer analytics firm can keep scraping public LinkedIn profiles, court says," Ars Technica, September 9, 2019, https://arstechnica.com/tech-policy/2019/09/web-scraping-doesnt-violate-anti-hacking-law-appeals-court-rules/.

[68] Adi Robertson, "Facebook's ad delivery could be inherently discriminatory, researchers say," The Verge, April 4, 2019, https://www.theverge.com/2019/4/4/18295190/facebook-ad-delivery-housing-job-race-gender-bias-study-northeastern-upturn. The Upturn study can be found here: https://arxiv.org/pdf/1904.02095.pdf.

[69] Devin Coldewey, "Facebook independent research commission, Social Science One, will share a petabyte of user interactions," TechCrunch July 11, 2018, https://techcrunch.com/2018/07/11/facebook-independent-research-commission-social-science-one-will-share-a-petabyte-of-user-data/; https://socialscience.one/.

[70] Shelby Brown, "Facebook opens data trove for academics to study its influence on elections: Researchers will get to parse Facebook ad data, the popularity of news items and URL data sets," CNET, April 29, 2019, https://www.cnet.com/news/facebook-opens-data-trove-for-academics-to-study-impact-on-elections/.

[71] Gary King and Nathaniel Persily, "First Grants Announced for Independent Research on Social Media's Impact on Democracy Using Facebook Data," Social Science One, April 28, 2019, https://socialscience.one/blog/first-grants-announced-independent-research-social-media%E2%80%99s-impact-democracy?admin_panel=1.

[72] Solomon Messing, Bogdan State, Chaya Nayak, Gary King, & Nate Persily, Facebook URL Shares, 2018, https://doi.org/10.7910/DVN/EIAACS, Harvard Dataverse, V2.

[73] Craig Silverman, "Exclusive: Funders Have Given Facebook A Deadline To Share Data With Researchers Or They're Pulling Out: Facebook has to provide key data by Sept. 30," BuzzFeed News, August 27, 2019, https://www.buzzfeednews.com/article/craigsilverman/funders-are-ready-to-pull-out-of-facebooks-academic-data.

[74] Solomon Messing, Twitter Thread on Social Science One Privacy Issues, August 23, 2019, https://twitter.com/SolomonMg/status/1164927631957143554?s=20.

[75] Ryan Williams and Manuel Blum, Presentation on k-Anonymity, Summer 2007, https://www.cs.cmu.edu/~jblocki/Slides/K-Anonymity.pdf#targetText=K%2DAnonymity%3A%20attributes%20are%20suppressed,a%20group%20of%20k%20individuals.

[76] Matthew Green, "What is Differential Privacy?" Cryptographic Engineering, June 15, 2016 https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/; see also Statement from Social Science Research Council President Alondra Nelson on the Social Media and Democracy Research Grants Program, Social Science Research Council, August 27, 2019, https://www.ssrc.org/programs/view/social-data-initiative/sdi-statement-august-2019/; Letter from funders to Social Science Research Council, August 27, 2019, https://ssrc-static.s3.amazonaws.com/sdi/resources/SMDRG_funder_letter_august_2019.pdf.

[77] For a list of areas where Facebook is active in research, see https://research.fb.com/research-areas/.

[78] Kate Klonick, Twitter thread, September 17, 2019, https://twitter.com/Klonick/status/1174001267330494473?s=20.

[79] Alex Abdo, "Facebook is shaping public discourse. We need to understand how: Social media platforms should lift restrictions impeding digital journalism and research," Knight First Amendment Institute at Columbia University, September 15, 2018, https://knightcolumbia.org/content/facebook-shaping-public-discourse-we-need-understand-how.

[80] "Facebook and Google: This is What an Effective Ad Archive API Looks Like," Mozilla, March 27, 2019, https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/.

[81] "Creating a French framework to make social media platforms more accountable: Acting in France with a European vision" (Final Mission Report on the Regulation of social networks – Facebook experiment, submitted to the French Secretary of State for Digital Affairs, May 2019), https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf.

[82] Institute for Strategic Dialogue, Online Harms White Paper Consultation, 2019, https://www.isdglobal.org/isd-publications/extracts-from-isds-submitted-response-to-the-uk-government-online-harms-white-paper/.

[83] Spandana Singh, "Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds," New America Foundation, October 21, 2019, https://www.newamerica.org/oti/reports/rising-through-ranks/.

[84] See Harold Feld, "The Case for the Digital Platform Act," Public Knowledge, May 7, 2019, https://www.publicknowledge.org/documents/the-case-for-the-digital-platform-act/; United Kingdom, Online Harms White Paper (Department for Digital, Media, Culture & Sport, April 30, 2019), available at https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper-executive-summary--2#contents.

[85] Mark MacCarthy, "A Consumer Protection Approach to Platform Content Moderation," in B. Petkova and T. Ojanen (eds), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries*, Edward Elgar, 2019 Forthcoming, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3408459.

[86] European Commission, Fourth intermediate results of the EU Code of Practice against disinformation, May 17, 2019, https://ec.europa.eu/digital-single-market/en/news/fourth-intermediate-results-eu-code-practice-against-disinformation.

[87] Institute for Strategic Dialogue, Online Harms White Paper Consultation, 2019, https://www.isdglobal.org/isd-publications/extracts-from-isds-submitted-response-to-the-uk-government-online-harms-white-paper/.

[88] Report Of The Facebook Data Transparency Advisory Group, Yale Law School, April 2019, https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf.

[89] The Santa Clara Principles on Transparency and Accountability in Content Moderation (May 7, 2018), https://santaclaraprinciples.org.

[90] Similar proposals for tiered access are found in Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015; Institute for Strategic Dialogue, Online Harms White Paper Consultation, 2019, https://www.isdglobal.org/isd-publications/extracts-from-isds-submitted-response-to-the-uk-government-online-harms-white-paper/; "Creating a French framework to make social media platforms more accountable: Acting in France with a European vision" (Final Mission Report on the Regulation of social networks – Facebook experiment, submitted to the French Secretary of State for Digital Affairs, May 2019), https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf.

[91] United Kingdom, Online Harms White Paper (Department for Digital, Media, Culture & Sport, April 30, 2019), paragraphs 29-30, available at https://www.gov.uk/government/consultations/online-harms-white-paper.

[92] S. ___, Augmenting Compatibility and Competition by Enabling Service Switching Act of 2019, Introduced by Senator Mark Warner, https://www.scribd.com/document/431507473/GOE19968.

[93] See Section 1(1) of the Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG), https://germanlawarchive.iuscomp.org/?p=1245.

[94] Some platforms for user-generated content might not be covered. For instance, Wikipedia is not a general interest forum for people to interact about topics of their lives. It does not amplify content. There is no share or like button. Things do not go viral on Wikipedia. Wikipedia also has no ads and doesn't aggregate user data that would be sold to advertisers. There's a good case that Wikipedia need not be covered by the transparency regulations described in this paper.

[95] S. ___, Augmenting Compatibility and Competition by Enabling Service Switching Act of 2019, Introduced by Senator Mark Warner, https://www.scribd.com/document/431507473/GOE19968.

[96] Reddit Privacy Policy, Effective June 8, 2018. Last Revised May 25, 2018, https://www.redditinc.com/policies/privacy-policy-may-25-2018.

[97] The evolution of the Wikimedia Foundation's terms of service is visible in its edit history: https://foundation.wikimedia.org/w/index.php?title=Terms_of_Use/en&action=history.

[98] GIFTC, Joint Tech Innovation: Hash Sharing Consortium, https://gifct.org/joint-tech-innovation/.

[99] Internet Watch Foundation, Hash List, https://www.iwf.org.uk/our-services/hash-list.

[100] The limitations of automated takedowns are well known and recognized by the social platforms themselves. But they are beyond the scope of this paper. For a good review, see Spandana Singh, Everything in Moderation: An Analysis of "How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content," July 22, 2019, https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/. See also Natasha Duarte, Emma Llanso, and Anna Loup, "Mixed Messages? The Limits of Automated Social Media Content Analysis" (Center for Democracy & Technology November 2017) http://proceedings.mlr.press/v81/duarte18a/duarte18a.pdf.

[101] Guy Rosen, "An Update on How We Are Doing at Enforcing Our Community Standards," Facebook, May 23, 2019, https://newsroom.fb.com/news/2019/05/enforcing-our-community-standards-3/.

[102] French Ambassador for Digital Affairs, Facebook's Ad Library Assessment, May 2019, https://disinfo.quaidorsay.fr/en/facebook-ads-library-assessment; Matthew Rosenberg, "Ad Tool Facebook Built to Fight Disinformation Doesn't Work as Advertised: The social network's new ad library is so flawed, researchers say, that it is effectively useless as a way to track political messaging," New York Times, July 25, 2019, https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html.

[103] "Facebook and Google: This is What an Effective Ad Archive API Looks Like," Mozilla, March 27, 2019, https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/.

[104] H.R.2592 - Honest Ads Act, Introduced by Rep. Derek Kilmer, May 8, 2019, https://www.congress.gov/bill/116th-congress/house-bill/2592; S.1356 - Honest Ads Act, Introduced by Introduced by Senator Amy Klobuchar, May 7, 2019, https://www.congress.gov/bill/116th-congress/senate-bill/1356.

[105] Katie Harbath and Sarah Schiff, "Updates to Ads About Social Issues, Elections or Politics in the US," Facebook, October 16, 2019, https://newsroom.fb.com/news/2019/08/updates-to-ads-about-social-issues-elections-or-politics-in-the-us/; Mark Zuckerberg, "The Internet needs new rules. Let's start in these four areas," Washington Post, March 30, 2019, https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html.

[106] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, & Harlan Yu *Accountable Algorithms*, 165 University of Pennsylvania Law Review 633 (2017), http://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3; Christian Sandvig, Kevin Hamilton, Karrie Karahalios, & Cedric Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms" Data and Discrimination: Converting Critical Concerns into Productive Inquiry, 2014, (Auditing Algorithms) available at http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf. A good example of what can be done without access to detailed information is the study by Upturn of Facebook's housing advertising practices, which can be found at https://arxiv.org/pdf/1904.02095.pdf.

[107] Karen Hao, "YouTube is experimenting with ways to make its algorithm even more addictive: Publicly, the platform says it's trying to do what it can to minimize the amplification of extreme content. But it's still looking for ways to keep users on the site," MIT Technology Review, September 27, 2019, https://www.technologyreview.com/s/614432/youtube-algorithm-gets-more-addictive/.

[108] For a good description of the issues arising in the credit context from new data and analytic models, see CFPB, Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, Federal Register/Vol. 82, No. 33/Tuesday, February 21, 2017, https://www.govinfo.gov/content/pkg/FR-2017-02-21/pdf/2017-03361.pdf.

[109] Institute for Strategic Dialogue, Online Harms White Paper Consultation, 2019, p. 9, https://www.isdglobal.org/isd-publications/extracts-from-isds-submitted-response-to-the-uk-government-online-harms-white-paper/.