TRANSATLANTIC WORKING GROUP

The Bellagio Session

Third session of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, November 13-16, 2019, hosted by the Rockefeller Foundation Bellagio Center, Italy

Co-Chairs Report No. 3

Observations and recommendations
—*Susan Ness and Marietje Schaake*

Working Papers

Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry
—*Mark MacCarthy*

Artificial Intelligence, Content Moderation, and Freedom of Expression
—*Emma Llansó, Joris van Hoboken, Paddy Leerssen, and Jaron Harambam*

Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective
—*Heidi Tworek, Ronan Ó Fathaigh, Lisanne Bruggeman, and Chris Tenove*

Co-Chairs Report No. 3: The Bellagio Session

Susan Ness, Annenberg Public Policy Center
Marietje Schaake, CyberPeace Institute

February 13, 2020

Introduction

The Transatlantic High Level Working Group on Content Moderation and Freedom of Expression (TWG) convened its third session as guests of the Rockefeller Foundation Center in Bellagio, Italy, from November 13-16, 2019.

Our first session, held in February at Ditchley Park in the United Kingdom, analyzed U.S. and European approaches to freedom of expression, and how these approaches could inform ongoing initiatives to address hate speech, terrorism, and other illegal speech online. Our second session, held in May at the Annenberg Beach House in Santa Monica, California, examined efforts to address online content that may not *per se* be illegal, but which may be considered "harmful." We discussed how maliciously deceptive material is virally spread with the intention of undermining informed debate that is essential in a democracy, and how that can be best addressed by focusing on the bad actors and dampening the virality of the messages (the behavior of the system) rather than the content.

At Bellagio, the TWG explored in detail three cross-cutting issues identified during our prior sessions: (1) transparency and accountability; (2) artificial intelligence and content moderation; and (3) dispute resolution mechanisms, including social media councils and e-courts. The group concluded that progress could be achieved on these issues from a multidisciplinary assessment, well-grounded in law, technology and business. The three research topics are intertwined.

As with prior sessions, draft briefing papers were circulated in advance of Bellagio and then deliberated at length under Chatham House Rule. Informed by the Bellagio discussions, the authors have revised their analyses. The final papers will be published shortly and posted on the IViR website. The opinions set forth in the papers remain those of the authors.

The TWG is a project of the Annenberg Public Policy Center (APPC) of the University of Pennsylvania in partnership with the Annenberg Foundation Trust at Sunnylands and the Institute for Information Law (IViR) at the University of Amsterdam.

TWG leadership transition

Marietje Schaake, president of the CyberPeace Institute and former Member of the European Parliament, has joined Susan Ness as co-chair of TWG, succeeding Nico van Eijk, who stepped down following his appointment as chairman of the CTIVD, the Netherlands Review Committee on the Intelligence and Security Services.

Preliminary observations and conclusions

As co-chairs, we offer the following preliminary observations and conclusions culled from the discussions in Bellagio. Members of the Transatlantic Working Group have reviewed our report and many of their comments are reflected in this co-chairs report.

Overarching themes

Our Bellagio session opened with a broader, philosophical conversation, which offered guidance throughout the session.

We briefly discussed an observation that speech has two divergent functions – discovery and deliberation – which cohabit in an age of information overabundance and distrust. The former pushes toward absolute freedom, the latter towards accountability. The internet has exploded with discovery, but has not helped very much on deliberation. How do we reconcile the two functions of speech to strengthen internet advancement of democracy?

How do we build sufficient transparency into the mechanisms by which business and democratic governments shape the public sphere to uphold rights and encourage healthy participation in that sphere? And when is human intervention essential?

We also discussed the "speech vs. reach" paradigm – the distinction between speech itself and the amplification of speech, either by paid advertising or by recommendation algorithms. What is the impact of amplification of speech beyond merely posting the speech itself? And do platforms have greater responsibility when they recommend content?

Reviewing our entire body of work, we agreed that the TWG must articulate an affirmative vision to enable democracy to remain resilient and to thrive. We were reminded that while Europe and the United States may differ in modest degree on the application of freedom of expression, we must think in broader terms about how authoritarian regimes such as Russia, China and others increasingly wield more control over the internet – both inside and outside their territorial boundaries.

As we address the rising volume and deepening impact of hate speech, violent extremism and viral deception online, we also must be prepared to tackle the growing sophistication of coordinated disinformation campaigns being launched now and in the future. It is a power battle, with those intending to do harm to democratic rights constantly improving their game. To ensure the resilience of democracy throughout the information ecosystem, collaboration between government, civil society and platforms/internet providers is essential. To lay the groundwork for such cooperation, a degree of trust between the parties must be fostered. As discussed below, transparency on the part of both platforms and government is key to building that trust.

We acknowledged the movement in many countries toward adopting a broad regulatory regime to address not just illegal and problematic speech online, but potentially other major concerns as well, such as privacy, copyright, and competition. Similarly, social media and other platform companies have begun to implement their own measures proactively to handle not only illegal but also harmful speech.

We encouraged greater transatlantic engagement in developing such frameworks to share best practices and to avoid unintended consequences – particularly with respect to freedom of expression, a cornerstone of our democratic systems – ever mindful that authoritarian regimes may cite western regulations to try to justify imposing harsher control over the online realm.

Finally, we experienced firsthand the value of transatlantic deliberations on issues of freedom of expression and human rights online, especially when enriched by participation from experts in law, technology and business. The TWG research and discussions have demonstrated concretely the benefit from both sides of the Atlantic coming together to learn from each other. We are deeply grateful that our work has been cited favorably in policy discussions around the globe.

Observations from the three research areas discussed at Bellagio

Emphasize and enforce platform transparency and accountability rather than regulating "legal but harmful" content

The "Transparency Requirements for Digital Social Media Platforms" paper outlines a transparency framework for those social media platforms that allow users to upload, share, and react to content. Most concerns regarding objectionable content arise in social media, where attempts to regulate can more easily infringe on the right to freedom of expression.

Instead of focusing on content regulation and mandatory removal of such content, the paper recommends a "balanced and clear legal structure for disclosure," expanding upon the French government proposal published in May 2019.

While the paper posits that a flexible government regulatory regime is the best approach for overseeing platform transparency and accountability, the industry is encouraged to adopt the transparency recommendations proactively and not wait for legislation to be enacted.

Social media platforms bring "communities" together under a platform-specific set of conduct rules – community standards and terms of service – which govern how a platform interacts with its users. Requiring a platform to clearly state its principles and conduct rules, disclose how these rules are being fairly and consistently enforced (including through automated curation), and offer a simple redress mechanism for users who believe their rights have been violated encourages healthier engagement online without violating freedom of expression.

Imposing and enforcing transparency and accountability requirements on internet platforms provides a less intrusive way to: (a) reduce the spread of "problematic" online content while protecting freedom of expression; (b) improve trust between platforms, government and the public; and (c) enable institutions to develop the capacity to draft flexible regulations in a dynamic environment. It also lessens the privatization of governance.

Improved transparency can also enable the forces of consumer choice, empowering users to protect themselves and to bring the pressure of public and political opinion to bear on social media companies. A focus on transparency enlists companies as partners in the effort to promote civil discourse. Strong transparency requirements also reassure the public and policy makers that platforms have policies and procedures designed to respect rights and address the challenges of hate speech, disinformation campaigns, and terrorist material.

- **Adopt a principle-based approach, flexibly applied**

Social media companies vary widely in business model, size and reach. A "one size fits all" regulation may be especially burdensome for smaller firms or companies that deliver specialized services to a limited segment of users. That said, social media platforms of all models and sizes should adopt community standards and terms of service and make them public in an accessible and user-friendly format. They should explain how they enforce such standards; publish procedures for complaints about standards violations as well as notification, review, and appeal processes; and report regularly on how they handled these cases. And, as discussed below, they should explain the criteria used in recommendation algorithms.

The community of users, as well as researchers and other outside interests (including the platforms' auditors), can help oversight bodies and the public ensure that the obligations the platforms undertake through terms of service and transparency requirements are fulfilled.

A transparency regime should provide different tiers of disclosure: for the public (outward); for oversight authorities and accredited researchers (inward); and, in the most protected cases, for regulatory authorities only. Greater standardization of data to be collected and published is essential, so that accredited researchers and regulators can better compare how well platforms are performing.

We note that many but not all platforms have made considerable progress in implementing transparency best practices.

- **Include algorithm-ordering and recommendation systems within transparency regimes**

For a transparency-based regulatory model to work, enforcement authorities must understand how platforms operate, including through the computer-based programs that amplify, rank, and moderate posted content (recommendation and prioritization algorithms). Information about these algorithms is needed to audit their role in disseminating and amplifying problematic content and to detect efforts to surreptitiously influence the formation of public opinion. It is not necessary to divulge the algorithm source code itself; rather, knowing the purpose and key factors can enable input/output testing to validate the algorithms' behavior.

Some contend that content referral algorithms recommend progressively violent or terrorist content in order to increase user engagement on the platform. These same algorithmic techniques could be used to recommend content promoting a particular political viewpoint or denigrating another. Although the referred content may be protected speech, the referral regime itself should be subject to transparency and accountability. As noted below, such transparency is essential when it concerns political speech. But in our highly polarized political world, we also must be wary of government using regulatory tools to achieve political ends.

- **Adopt clear transparency rules for political advertising**

Platforms should provide robust disclosure surrounding political advertising and the use of platforms by politicians, including verified accounts. For example, legislation introduced in the U.S. Congress (the Honest Ads Act), like the EU's Code of Practice on Disinformation, would require large platforms to maintain a searchable public file with a copy of the political ad, disclosure of the sponsor, the amount spent, the targeted audience, and number of views. The platform also would have to use reasonable efforts to ensure that foreigners are not purchasing political ads to influence American elections. Such transparency requirements enhance rather than harm freedom of expression.

- **Work within the internet's global reach …**

A principled and flexible transparency-based approach to online content moderation is better suited to the internet's global reach. Most platforms, regardless of size or model, are accessible globally, but the various legal protections offered for users across jurisdictions raise the possibility of conflict of laws. While transparency requirements may vary between jurisdictions, the tiered approach recommended in the briefing paper should satisfy most regulatory requirements.

- **… and within the transatlantic community**

Because even an enforceable transparency-based regulatory model may be implemented differently across jurisdictions, there is a compelling case for transatlantic collaboration on the approach, given the enormous flood of internet traffic across the Atlantic and our shared commitment to democracy and freedom of expression as well as universal human rights.

TWG members noted many different avenues for such discussions, including bilateral contacts between legislators and agencies, the U.S.-EU Information Society Dialogue and the OECD. All should be encouraged, together with multistakeholder engagement.

Understand the benefits and limitations of artificial intelligence

Technology is not neutral, as those who build and program it inevitably bake in certain values. Developments in computing like artificial intelligence (AI) and machine learning can serve as both a positive and a negative force on human rights and fundamental freedoms. Such tools, including simpler forms of automation and algorithmic systems, can help in identifying at massive scale some forms of illegal content, such as child pornography or terrorist propaganda. And they have been used successfully in countless content referral situations, such as recipe recommendations. But they are not a silver bullet. They are only as effective as the datasets that train them (bias in, bias out) and the suitability of the task to which they are assigned (i.e., search engines versus social media ordering).

Data inputs used to train the programs may be flawed, biased, and incomplete, especially when dealing with smaller datasets involving non-Western cultures, communities and languages. Intended and unintended consequences may vary greatly. Small variations can disrupt patterns, and AI often has difficulty assessing context and nuance. As a result, regulations that explicitly require or push platforms to over-deploy these techniques risk creating many false positives against legitimate speech in order to minimize the amount of "harmful" content remaining online.

For smaller platforms with fewer resources to create, maintain and update programs to screen content, the problems of misidentification or failure to identify are even more acute, potentially leading to greater liability (i.e., for failure to catch copyright violations.) Using the datasets of larger platforms could bias in favor of Western or Chinese outcomes, or could violate privacy rules. In sum, despite great computing power, automation systems are not reliable, and are not ready to shoulder without human intervention the full responsibility for content moderation.

Finally, tasking private companies to address "harmful" content to safeguard the public interest raises serious governance issues.

These technological limitations and pitfalls are described in detail in the TWG paper on "Artificial Intelligence, Content Moderation and Freedom of Expression." The paper serves as a much-needed

primer for policy makers on both sides of the Atlantic to clarify the structure and uses of tools collectively known as -- or mistaken for -- artificial intelligence. It also reflects on the need for new freedom of expression safeguards tailored to such automated forms of speech governance.

- **Adopt consumer safeguards for use of AI recommendation/ranking functions**

Powerful automated systems also are used for content dissemination through recommendation/prioritization functions. These can be driven by organic sharing by individuals, or they can be inorganically shaped to promote certain content feeds in response to expressed or inferred user interest or other amplification signals, including paid promotions.

Such prioritization programs – whether in social media, news feeds, retail platforms, or search engines – are essential to the internet because they make an otherwise overwhelming amount of information manageable.

But even as these programs can benefit users, they can mislead them. For example, search results can be tainted by "data voids." These are search engine queries that turn up few or no results, often concurrent with a major event unfolding. Manipulators can exploit these data voids by rapidly and repeatedly linking these queries to problematic content, such as hate symbols, conspiratorial content, or other disinformation, to fill the void. The result is compounded by "autofill" or "autoplay" technology promoting "trending topics" that then are amplified by mainstream media. To avoid manipulation during major events, some platforms have locked pages, and have privileged "verifiability" over "truthfulness."

Efforts to train prioritization programs by boosting "authentic" reporting and/or down-grading or demoting information that does not meet fact-checking standards are helpful but insufficient. Some users claim that these mechanisms are biased against their point of view. These concerns are heightened by the lack of insight into how prioritization programs work.

Platforms that deploy these systems should provide greater transparency about the use of these tools and the consequences for consumers. Review of such systems should be included in any transparency oversight regime. Enabling more transparency, explicit user choices, and control over material they see – coupled with consumer education – should help to curtail abuse.

A flexible transparency-based approach can enable accountability by allowing regulatory authorities and vetted researchers reasonable access into both the design of the algorithms and their operational effects, as well as better inform the companies about unintended effects.

- **Use caution when addressing political content and referral algorithms**

During political election seasons, there is heightened apprehension over the use of algorithmic referral systems and/or paid political advertising to manipulate surreptitiously what the internet user/voter sees concerning a particular candidate or policy issue. This matter both affects freedom of expression as well as the ability to have an informed electorate – which is essential to democracy. During the 2016 U.S. presidential election, microtargeting was extensively deployed below the radar, based upon political preferences inferred from large personal datasets. Some people received microtargeted ads that were crafted to increase polarization or to reduce voter turnout.

As noted, legislation has been introduced in the U.S. Congress for robust disclosure and labelling of online political and issue advertising, the funding source, and the real party in interest, including the

number and selection criteria for the people targeted. This echoes legislation and regulation already in effect in the EU and a number of European countries.

In addition, major social media platforms have responded by adopting different approaches to address political advertising. At least one has ceased accepting political advertisements, while others will limit microtargeting to certain categories. Still others will not interfere with candidate statements or ads, supporting the principle that the public has the right to hear directly from candidates without corporate intervention.

Political communication should have special protected status. Consumers have a right to know how they are being targeted, and by whom. Legislation is needed to set transparency rules for political advertising and microtargeting. Reasonable limits on microtargeting by political campaigns would not diminish freedom of expression.

Platforms should maintain a comprehensive archive of political advertising so that vetted researchers under strict privacy rules can analyze whether voters are being manipulated (i.e., are subjected to bot-driven campaigns or disinformation.) Researcher access to these archives will also contribute to better informed policy decisions.

Establish efficient and effective dispute resolution systems for social media platforms

Decisions by governments, companies or even online communities to remove, promote, demote, or demonetize content created and uploaded by individuals, as well as refusals to remove content, immediately raise concerns about the right to freedom of expression. This is most immediately obvious when governments constrain the freedom of expression – a step that should be done only through considered rule of law protections and democratic processes.

Especially in the United States, but also in Europe, companies and communities have freedom of expression rights of their own to set and enforce standards for permissible conduct while respecting the law. But users whose content is removed or downgraded by social media companies under their terms of service/community guidelines should have a right to contest and appeal such decisions, both with the company or community and, ultimately, through a redress process when the user believes that the platform itself is violating the contractual rights embodied in the community standards and terms of service. Given the increasing use of automated takedown systems, that possibility grows.

More problematic still is when governments outsource censorship by pressuring platforms to remove "objectionable" but not illegal content without the normal judicial process required under international human rights laws and in democratic systems of governance. Democratic societies should not privatize the protection of the freedom of expression. This right should be protected through independent judicial systems.

The TWG paper "Dispute Resolution and Content Moderation: fair, accountable, independent, transparent, and effective" asserts that social media councils – whether at the global, national or corporate level – could provide independent guidance on content moderation standards and procedures, and could even be used to adjudicate disputes. For cases that specifically involve potential violations of human rights, including but not limited to freedom of expression, a form of online judicial determination, or internet court (e-court), should be considered.

- **Establish social media councils for policy advice or dispute resolution under criteria of fairness, accountability, independence, transparency and effectiveness**

Many social media companies have internal procedures to enable appeals about content that may have been wrongfully removed, or where other users cite offensive content that they believe violates community standards and has not been removed. Generally, companies alone determine the mechanisms for review and render the decisions. While welcome, such internal mechanisms do not meet the essential rule of law standards of a good dispute settlement system: fairness, accountability, independence, transparency, and effectiveness (FAITE).

The scale of the problem is enormous. For example, in the Facebook Transparency Report for the third quarter 2019, Facebook took down over 7 million pieces of content under its own global hate speech rules. Users appealed 1.4 million of these takedowns, and the company restored just 12% with no further process cited. Facebook employs around 30,000 reviewers across the globe, although most initial screening is performed by algorithms with limited human intervention. Smaller firms, however, may not have the staffing or financial resources to replicate these review mechanisms. They also generally have less reach and impact than do the major platforms (although they may be the dominant player in a country with a smaller population).

A high-level, strictly independent body to make consequential policy recommendations or to review selected appeals from moderation decisions could go a long way toward improving the level of trust between platforms and the public. As detailed in the TWG paper, there are a wide range of organizational structures and precedents to consider, with the format, jurisdiction, makeup, member selection, standards, and scope of work subject to debate. At this juncture, experimentation by platforms and multistakeholder groups will provide invaluable data points to guide future structural decisions.

- **Consider establishing an e-court system for rapid determination of fundamental rights violations**

For appeals predicated on fundamental rights, the concept of an e-court has considerable merit. As discussed in Bellagio and Santa Monica, an e-court system enhances legitimacy of the process through the rule of law, independence, and impartiality from the parties.

It would provide an online procedure for users to challenge content moderation decisions made by social media companies. Specially trained magistrates would rule quickly on the simple question of whether the removal or refusal to remove was consistent with legally cognizable rights. The regular publication of case-law compilations would create a body of precedent. The degree of scalability is yet to be determined.

Europe, Canada and the United States have various models of expedited resolution systems that can inform the design of an e-court system. The e-courts would be funded by government and/or by contributions from platforms. Such an expedited judicial review procedure could be complemented by other online mediation and arbitration procedures that meet the FAITE standard.

Next Steps

The three Bellagio papers will be widely circulated to policy makers and stakeholders. We encourage reader feedback and discussion. The TWG intends to hold several roundtables with policy makers and stakeholders to further refine our views. We plan to issue a final report in the spring of 2020, informed by that feedback, with launch events in both Europe and North America.

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/.

TRANSATLANTIC WORKING GROUP

Artificial Intelligence, Content Moderation, and Freedom of Expression†

Emma Llansó, Center for Democracy and Technology[1]
Joris van Hoboken, Institute for Information Law, Vrije Universiteit Brussels[2]
Paddy Leerssen, Institute for Information Law[3]
Jaron Harambam, Institute for Information Law, University of Amsterdam[4]

February 26, 2020

Contents

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

Introduction

As governments, companies, and people around the world grapple with the challenges of hate speech, terrorist propaganda, and disinformation online, "artificial intelligence" (AI) is often proposed as a key tool for identifying and filtering out problematic content. "AI," however, is not a simple solution or a single type of technology; in policy discussions, it has become a shorthand for an ever-changing suite of techniques for automated detection and analysis of content. Various forms of AI and automation are used in ranking and recommendation systems that curate the massive amounts of content available online. The use of these technologies raises significant questions about the influence of AI on our information environment and, ultimately, on our rights to freedom of expression and access to information.

What follows is a compact position paper, a first version of which was written for the Bellagio, Italy, session of the Transatlantic Working Group (TWG), Nov. 12-16, 2019. It discusses the interface of AI/automation and freedom of expression, focusing on two main areas.

Part I focuses on **content moderation** and the use of automated systems for detecting and evaluating content at scale. Part II focuses on **content curation** and questions about the role of recommendation algorithms in amplifying hate speech, violent extremism, and disinformation.

For both content moderation and content curation, the paper explores the use of AI and other forms of automation. In particular, it focuses on their use in the fight against hate speech, violent extremism, and disinformation. On that basis, we reflect on the need for new freedom of expression safeguards tailored to these new automated forms of speech governance.

This paper and its recommendations take into account the transatlantic nature of the TWG and are therefore intended to speak to both U.S. as well as EU legal contexts. The authors are grateful for the discussions and observations produced by participants in the TWG meeting at Bellagio, many of which have been incorporated into this report and the recommendations that follow. The text of the paper remains the sole responsibility of the authors.

Key Recommendations

1. Speech governance debates should not focus exclusively on AI technology as such, but take into account a broader range of automating technologies and processes, including simpler forms of automation and algorithmic systems.

2. Automation in content moderation should not be mandated in law because the state of the art is neither reliable nor effective.

3. Entities that use automation in content moderation should provide greater transparency about their use of these tools and the consequences they have for users' speech, privacy, and access to information.

    a. Priority should be paid to enabling availability of research data while respecting users' privacy and data protection rights, and in particular to resolving the questions of

whether or how data protection regimes in the U.S. and Europe constrain the sharing of data for research.

4. Policy makers should resist simplistic narratives that lay the blame for harmful content online exclusively with "algorithms." Instead, policy makers should recognize the role of users and communities in creating and enabling harmful content. Solutions to policy challenges such as hate speech, terrorist propaganda, and disinformation will necessarily be multifaceted.

5. Regulating ranking algorithms raises free speech risks comparable to outright removal of content and should be treated with comparable safeguards. Content regulation via ranking decisions is particularly problematic due to the lack of transparency in this space and the potential for far-reaching and unintended consequences for online discourse.

6. There is no such thing as a "neutral" recommendation algorithm and policy makers should therefore avoid simplistic mandates for "neutrality" or "non-discrimination."

7. Potentially fruitful measures to address harmful content in recommendation systems include:

    a. Enhancing user agency by providing the option to choose between different approaches to ranking.

    b. Media literacy and awareness programs.

    c. Enhancing public transparency about the design and performance of recommendation systems.

Part I: Automation in Content Moderation

Introduction

Enormous amounts of content are uploaded and circulated on the Internet every day, far outpacing any intermediary's ability to have humans analyze content before it is uploaded.[5] Many companies and governments are turning to automated processes to assist in detection and analysis of problematic content, including disinformation, hate speech, and terrorist propaganda.

"AI in content moderation" is a very broad concept: In one sense, there is very little true "artificial intelligence" in use in content moderation systems.[6] As it is typically used in law and policy conversations, however, AI in content moderation can refer to the use of a variety of automated processes at different phases of content moderation. These processes may be as simple as keyword filters, but they may also rely on machine learning and can include a number of tools and techniques. In this section we will examine a variety of machine learning processes and discuss how they are used in different phases of content moderation.

Content moderation on a site or service with a global user base requires the application of a set of rules or standards for speech that may occur in many formats (such as text, images, video, or audio), on any subject, in potentially any language or dialect, from thousands or millions of cultural contexts. As a recent report from Cambridge Consultants noted, content moderation raises a number of challenges, "such as ambiguity within a certain piece of content, the change of meaning within a post

when context is considered and the potential for bias in the moderator. AI moderation has to contend with the same set of contextual challenges, along with a number of AI-specific technological challenges."[7]

In content moderation, automation can be used in the related phases of proactive **detection** of potentially problematic content and the automated **evaluation** and enforcement of a decision to remove, tag/label, demonetize, demote, or prioritize content. Machine learning tools can also be used to automate the **generation** of content and accounts. Understanding these capabilities for creating new material can also aid in the detection of it, including tools for the evaluation of manipulated imagery, such as "deepfakes."[8]

Key concepts in machine learning

Supervised learning involves training a model based on a labeled or annotated training dataset. For example, a research team that is building a tool to identify hate speech on a social media platform could label a corpus of posts as "hate speech" and "not hate speech" (or as "racist," "homophobic," "anti-immigrant," and "not hate speech," or any other set of labels). These labels assist the machine learning model in identifying features of the data that help to distinguish the various categories of posts from one another. Supervised learning models can be effective at developing classifiers to distinguish or categorize different inputs, but they can be resource-intensive to train, as they require substantial quantities of hand-labeled inputs. Further, as discussed below, the processes of generating the training dataset and having one or more people label it can introduce biases and errors into the model.

Unsupervised learning involves training a model based on an unlabeled dataset; the model learns to identify underlying patterns and features within the data. Unsupervised learning can, for example, be used to develop a corpus of word pairs that often occur together in a text. These pairs can then serve as the "labeled" training data for a tool like Word2Vec that assesses the relationships among words called "word embeddings." [9] Word embeddings can be a powerful way to automatically parse text, but such tools can also inadvertently "learn" associations in the underlying text that reinforce cultural biases.[10]

There are a number of different types of algorithms that are frequently used in machine learning. For instance, supervised learning often uses linear regression or decision tree algorithms, whereas unsupervised learning uses K-means clustering or Apriori association rule learning.

Discriminative algorithms are those that apply a classifier to determine whether an input should be labeled as fitting into a particular category (e.g., spam or not spam). *Generative algorithms*, by contrast, start from a particular label and predict what features or characteristics the "input" should have.

Mathematical *models* are what encode the results of the machine learning done on the dataset. A *neural network* is a layered machine learning model akin to the connections among neurons in the human brain. *Convolutional neural networks* (CNN) are a type of neural network commonly used to evaluate image data.[11] *Recurrent neural networks* (RNN) are neural networks that incorporate outputs from the system into subsequent runs and are thus better able to process sequences of information.[12]

Capabilities and limitations of automated content analysis

Analysis of text

Automated blocking and removal of online text predates sophisticated machine learning techniques: some of the earliest automated content moderation systems relied on keyword filtering to block posts or access to websites that included certain words and phrases. Keyword filtering is notoriously overbroad and underinclusive, blocking words regardless of their context or meaning and failing to filter content not specified on the list of prohibited terms.

Natural language processing (NLP) is a field of study that seeks to enable computers to parse text in a more comprehensive way, closer to the way that a human would understand the text.[13] NLP tools can be trained to predict whether a text is expressing a positive or negative emotion (sentiment analysis) and to classify it as belonging or not belonging to some category (such as the hate speech classifier described above). NLP tools are often trained on text that has been stripped of features such as URLs, usernames, and multilingual communication, but some researchers are beginning to experiment with incorporating information from emojis into sentiment-analysis tools.[14]

A well-known example of an NLP tool is Google/Jigsaw's Perspective API, an open-source toolkit that allows website operators, researchers, and others to use Perspective's machine learning models to evaluate the "toxicity" of a post or comment.[15] Perspective provides a good illustration of both the capabilities and limitations of a sophisticated NLP tool. It has been used for a variety of applications,[16] including as a tool that comment moderation systems use to warn users that they may be posting a "toxic" comment and to give them the opportunity to revise their comment.[17] But Perspective, and the concept of evaluating "toxicity" of comments, is far from perfect; soon after the Perspective API was launched, researchers began exploring ways to "deceive" the tool and express negativity that slipped under the radar,[18] and researchers continue to identify bias in the tool, such as misclassification that disproportionately affects different racial groups.[19] The Conversation AI research team behind Perspective cautions, "*We do not recommend using the API as a tool for automated moderation*: the models make too many errors."[20]

Machine learning models can also be used in the generation of text. Earlier this year, the OpenAI research team announced the public release of their GPT-2 language model, a predictive text tool that was trained on a dataset of eight million web pages.[21] The text generated by the GPT-2 model can be fairly complex and, according to the researchers, regularly outperforms the previous state-of-the-art models for text generation.[22] The research team made a widely publicized decision to release only a smaller, less capable model for public scrutiny and use, "[d]ue to concerns about large language models being used to generate deceptive, biased, or abusive language at scale."[23] Six months later, it released a larger version of the model along with a paper discussing the lessons it had learned through this "staged release" strategy which has allowed OpenAI and other researchers to more fully consider the implications and potential malicious uses of this technology.[24] (Meanwhile, two graduate student researchers at Brown University attempted to replicate OpenAI's full GPT-2 model and released their version publicly in August 2019.)[25]

Analysis of images

Automated image-detection and -identification tools can range from fairly simple systems designed to detect previously identified content to more complex tools designed to discover features of novel content. *Hash values* are unique numerical values that are generated by running a specific algorithm on a file; the *hash function* calculates the numerical value based on characteristics of the file, and can be thought of as generating a specific "digital fingerprint" for that file. A system can run the same hash function on novel/newly uploaded content and detect whether the novel content matches the hash value of previously identified content.

For simple hashing, the characteristics that the function evaluates can include things like dimensions of the image and specific color values of pixels; changing any of these characteristics, however, can completely change the hash of the altered file, which makes simpler hashing functions easy to circumvent if the goal is to evade detection.[26] A more sophisticated approach, *perceptual hashing*,[27] can be more resilient against circumvention by calculating hashes based on relationships among pixels and accounting for minor variations in the resulting hash.[28] Microsoft's PhotoDNA tool, for example, converts an image to black and white, resizes it to a standard size, divides the image into a grid, and calculates the hash based on the intensity gradient of each black-and-white square.[29] These transformations of the image, as part of the hash function, help to counteract minute changes to the image that could stymie simpler hash functions.

Other approaches to image analysis, including machine learning methods and techniques from the field of *computer vision*, work to detect the presence of specific elements or features in an image, such as symbols or logos, weapons, or nudity.[30] Tools that are designed to detect pre-identified images, such as a specific symbol or logo, need to be able to identify variations of that symbol in different lighting conditions, resolutions, and rotation/skew. *Optical character recognition* tools can identify text in an image and convert it into machine-readable formats, which is an essential step to using natural language processing to evaluate the meaning of text.

Tools can also be designed to classify whether an image contains a feature such as nudity. One approach to detecting nudity in an image has been to analyze the proportion of pixels in an image that fall into a specific color range that has been pre-identified as representing skin color. This kind of tool is vulnerable to misclassification of underrepresented skin tones and of objects or scenes with the same color palette as the training data (for example, deserts[31]). More involved machine learning tools will use skin-tone detection as a component of their analysis, along with other image-parsing processes to detect, for example, the presence of faces[32] and distribution of body parts. Such tools then use this underlying information about the components of an image to generate a classifier to identify likely nudity or sexual activity. Even more complex tools, however, can erroneously classify content to a significant degree when they are used on mass-scale, highly variable datasets like social media postings, as Tumblr discovered when it implemented its ban on nudity and sexual content, which yielded a wide variety of false positive results.[33]

Other deep-learning methods enable techniques such as scene understanding, which not only identifies the discrete features/likely objects in an image but analyzes them in the context of their relationship with the other objects in the image.[34]

The field of image-generation is also rapidly developing. For example, *Generative Adversarial Networks* (GAN) use generative algorithms to create new data to test and refine a machine learning classifier. GANs can be useful for training a machine learning model in how to detect a manipulated image or video.[35] Generative algorithms can also be used to deceive machine learning tools by manipulating images so that they look essentially unchanged to the human eye but display mathematical features that the classifier will understand to be something else entirely.[36] Researchers have demonstrated effective adversarial techniques even against black-box networks, i.e., networks for which the attackers have no specific knowledge of the model or the training data that generated it.[37]

Image generation also notably includes the area of "deepfakes," composite videos and images created on the basis of real footage that portray fictional statements and actions. A common form of deepfake is face-swapping, where the face of one individual is superimposed on the body of another. For example, in a video aimed at warning people not to believe everything they see on the Internet, filmmaker Jordan Peele created a deepfake video where his words appear to be spoken by President Obama.[38] To accomplish this, an *autoencoder* is used to analyze a large volume of images of a person to create a detailed mathematical map of the features of an individual's face (encoding) and to develop a process for turning these features back into the image of the individual's face (decoding). Once the autoencoders are trained, the encoded data of one person's face can be translated back into an image by the decoding process for another person's face, essentially preserving the features of the first person but placing them in the context of the second person's face.[39] A GAN can then be used to evaluate and iterate on the produced images and video and develop a more refined outcome.[40] Deepfakes can threaten rights to privacy and dignity, as in the case of involuntary pornography,[41] and may be used in disinformation campaigns. However, experts caution that the expense and time required in creating deepfakes will likely limit their use in disinformation campaigns, especially when it is easier to create misleading or recontextualized information to achieve similar results.[42]

There are a variety of other forms of automated content analysis that can be relevant in content moderation, including video and audio analysis techniques and efforts to detect bot networks/accounts.[43]

Technical limitations of automation in content moderation

Different technical approaches to automated detection and analysis of user-generated content will have limitations specific to their design – for example, a tool designed to detect "toxic" comments in one language will have difficulty parsing multilingual text.[44] This section summarizes major technical and design limitations to automated content detection and analysis; Section 1.3 examines the broader limitations of these tools and the risks for freedom of expression when they are incorporated into content moderation systems at mass scale.

The importance of context: Whether a particular post amounts to a violation of law or content policy often depends on context that the machine learning tool does not use in its analysis. Some context could be incorporated into a machine learning tool's analysis, such as the identity of the speaker or the relationship between sender and receiver of a message, but these come with significant tradeoffs for privacy. Other context, such as historical, political, and cultural context, are much more difficult for a tool to be trained to detect.

Lack of representative, well-annotated datasets to use for training: Machine learning tools develop their ability to identify and distinguish different kinds of content based on the datasets they are trained on. Many tools are trained on labeled datasets that are already publicly available; if these datasets do not include examples of speech in different languages and from different groups or communities, the resulting tools will not be equipped to parse these groups' communication.

Annotation for supervised learning can introduce bias: The process of labeling a dataset for supervised learning typically requires the involvement of multiple human beings to evaluate examples and select the appropriate label, or to evaluate an automatically applied label. *Intercoder reliability* is an important measure of how consistently different humans involved in labeling a dataset perform this task. Low intercoder reliability means that the humans applying the label do not agree among themselves what content merits the label of, for example, "hate speech" or "spam."

Need for flexible, dynamic models: Human communication patterns can change quickly, and speakers who are blocked by automated filters often have extra incentive to figure out how to circumvent the filter. Static machine learning models will quickly become outdated and unable to correctly classify users' communications.

Domain specificity: Natural language processing tools perform best in environments that closely match the data they were trained on. It is difficult to develop tools that work well across a variety of sites, languages, cultures, interest groups, and subject matter.

Significant risks of bias against underrepresented speakers: Applying a tool to a domain or group of speakers who do not closely match the groups represented in the training data can lead to erroneous classifications that disproportionately affect underrepresented groups.[45]

Resource limitations in energy, data, and processing power: Geoffrey Hinton, "one of the forefathers of modern deep learning," argues that we may have hit the limits of what existing machine learning techniques can do. "Hinton points out that CNNs [convolutional neural networks] are highly inefficient at learning features: neural networks require huge amounts of memory and computing power, and massive amounts of data, and still struggle with translational and rotational changes of objects."[46] Wide-scale application of ultra-sophisticated machine learning models for content analysis may be too resource- and energy-intensive to be sustainable.

Analysis: Freedom of expression threats and safeguards

Beyond the technical limitations of any particular tool, the use of automation in content moderation systems raises distinct challenges for freedom of expression and access to information online. There is a growing body of literature on the human rights implications of the use of automation by online services: The UN Special Rapporteur on freedom of expression, David Kaye, has offered analysis, conclusions, and recommendations on artificial intelligence in a number of recent reports.[47] The Council of Europe has provided several reports, studies, and recommendations that touch on the topic and is in the process of finalizing a new recommendation on the human rights impacts of algorithmic systems.[48]

It is important to note that the application of artificial intelligence in the online media environment can have both positive and negative implications for individuals' right to freedom of expression. First,

AI is at the core of the services that are central to effectuating people's rights to express themselves and to access information. Search engines, social media, and other internet services deploy various complex and adaptive information-processing technologies at the core of their operations. Without these technologies, these services, and the value they provide to people in expressing themselves and accessing information, would simply not be possible. This is not to say that relevant services are always doing a good job in supporting freedom of expression. There are significant concerns about how well current services serve our democracies and respect people's right to freedom of expression. Even so, the deployment of these advanced data-processing operations will remain central to our media and communications ecosystem.

Second, algorithmic systems have become a necessary tool to defend freedom of expression and the values underlying it. AI not only powers complex service operations, it is increasingly necessary to create the conditions for a robust and vibrant democratic exchange on online platforms. To make a simple analogy: Without AI, our media would increasingly feel like an email inbox without a spam filter.

Third, such filtering systems, including content moderation and recommendation systems (see Section II), raise their own set of freedom of expression concerns. Their application can raise issues of bias and discrimination, private due process, and surveillance issues that our current legal frameworks have not fully addressed. When combined with regulatory pressure on platforms to tackle issues such as disinformation, terrorism content, and hate speech, the application of these tools raises privatized censorship concerns, as well as questions about prior restraint, and due process.

AI tools and risks to freedom of expression

There are a number of recognized issues with the application of algorithmic systems and automation for these purposes:

False positives and false negatives: The use of algorithmic systems for detecting particular types of speech and activity will always have so-called false positives (something is wrongly classified as objectionable) and negatives (the automated tool misses something that should have been classified as objectionable). From a freedom of expression perspective, the implications of false positives and negatives depend on the goals of the tool that is used. If the tools are used to identify and demote or remove content, or single it out along with the relevant content creators for further scrutiny, false positives risk significant burdens on individuals' right to freedom of expression. False negatives, on the other hand, can result in a failure to address hate speech, harassment, and other objectionable content that may create a chilling effect on some individuals' and groups' willingness to participate online.

Potential bias and algorithmic discrimination: Algorithmic systems have the potential to perform badly on data related to underrepresented groups, including racial and ethnic minorities, non-dominant languages, and/or political leanings. This is due both to the lack of data and to the possibility of biased training datasets: If data are influenced by real-world biases and inequalities, then the models trained on these data may come to reflect or amplify these inequalities. This can result in

serious risks to freedom of expression for communities and individuals, potentially including illegitimate silencing of their expression and failure to address harms to their communities.

Large-scale processing of user data and profiling: Algorithmic systems will typically rely on the large-scale processing of user data to develop and apply the tools. These systems may also involve the additional profiling of users in view of the risk that such users engage in activity that may warrant additional scrutiny and/or risks from a content moderation perspective. In this way, the growing reliance on algorithmic systems further encourages the collection and processing of personal data, which pose additional risks to the rights to privacy and freedom of expression.

Presumption of the appropriateness of prior censorship: Automatically pre-judging content and prohibiting it from being posted is the very definition of prior restraint/censorship. While pre-screening content to limit the spread of malware, child abuse material, and spam has been broadly accepted as a positive use of automation, we must be cautious about applying that logic to other types of speech.

Inadequate oversight and lack of due process: Algorithmic systems may be applied as a quick fix for the complex task of judging whether particular content or activity warrants restrictive actions (which can range from flagging, demoting, demonetizing or removing it, to actions taken against account holders). Without proper complaint, review, and appeal procedures as well as oversight, these actions may violate freedom of expression rights, due, for instance, to the issue of false positives and negatives identified above. In particular, this raises due process issues. And given the difficulty of explaining and documenting complex machine learning systems, it may be even more difficult to create transparency and monitoring here than for other types of content moderation. Fundamentally, there is a tension between the evaluation of AI tools from a statistical perspective (how well are they performing overall) and their evaluation on a case-by-case basis, which is the predominant mode of evaluation from a fundamental rights perspective.

Need for redress and accountability: Automation in content moderation challenges preexisting structures and frameworks for making determinations about speech, given the enormous scale of speech that is being evaluated. What do remedy – and accountability – look like at scale?

The role of platforms in communications governance

Online platforms facilitate and shape the power of the speech of others.[49] They can act (or perhaps better, be asked or forced to act) as control points in tackling the proliferation of illegal and different forms of harmful speech and activity. What their role should be, and by what standards they should be judged, remains a topic of intense debate.

In part because of their critical role in facilitating speech, online platforms are also at the center of discussions about objectionable content. Not surprisingly, online platforms are now central sites for the development of automated content moderation systems and solutions. To develop these systems, larger platforms such as Google and Facebook can leverage world-leading expertise in machine learning and unrivaled financial resources in combination with large datasets and internet-scale user activity. The role of these private sector actors in developing content moderation tools may result in

a lack of transparency and accountability with respect to their functions and effects, including on freedom of expression. In addition, the development and implementation of these technologies will be informed by specific commercial motivations that are connected to the platforms' business models. Any freedom of expression safeguards with respect to the use of algorithmic systems in content moderation will have to address the central role of platforms in the development and application of these tools in real-world settings.

Recommendations

The human rights framework

The implications of artificial intelligence in media and communications for human rights have already been explored. As the UN special rapporteur on freedom of expression concluded in his thematic report to the General Assembly,

> AI tools, like all technologies, must be designed, developed and deployed so as to be consistent with the obligations of States and the responsibilities of private actors under international human rights law. Human rights law imposes on States both negative obligations to refrain from implementing measures that interfere with the exercise of freedom of opinion and expression and positive obligations to promote the rights to freedom of opinion and expression and to protect their exercise.[50]

Further development and deployment of algorithmic systems will continue to raise new issues under the existing human rights framework that should be addressed. One important aspect to take into account is the positive obligation for governments to encourage pluralism and diversity. The online information ecosystem raises a variety of questions with regard to pluralism and diversity: On the one hand, many speakers and perspectives have been able to attain a broader audience than ever before. But on the other hand, the dominance of a few major online advertising providers has radically changed the business model, and fiscal stability, of news and media publishers around the world. Moreover, the application of algorithmic systems for content moderation raises distinct questions of pluralism, as these systems impose a particular set of content restrictions at potentially enormous scale. Policy makers should evaluate the risk to pluralism and diversity that may come from automated content moderation on massive platforms. Policy makers and platform operators alike should consider specific safeguards designed to achieve pluralism and diversity. Such safeguards could range from clear, minimally restrictive default settings combined with greater user controls or Application Programming Interfaces (APIs) that enable users to implement a third-party's content moderation rules on the content they view on a platform.

Regulation of platforms vs. other forms of regulation

Increasingly, regulation is targeting platforms, due to their central role and significant power in the media and communications landscape. Many of the self- and co-regulatory initiatives in this area are also coming from platforms. This can leave other (and existing) forms of regulation to address the fundamental rights implications of algorithmic systems for content moderation unexplored. In Europe, the General Data Protection Regulation (GDPR), which is broadly applicable to the processing of user data in the online environment, provides an important baseline for the protection

of fundamental rights in data-driven power dynamics between platforms and their users. It sets limits to profiling activities, imposes transparency and accountability requirements, and grants relevant rights to individuals. Non-discrimination law has an important role to play in addressing bias and discriminatory impacts of algorithmic systems. Antitrust law should help to prevent undue concentrations of power over media and communications. It's crucial to acknowledge the importance of a broad set of regulatory frameworks that provide important baseline conditions for the effective exercise of freedom of expression online.

Intermediary liability

Intermediary liability frameworks, discussed in a separate TWG discussion paper,[51] have always had a strong link to the state-of-the-art of technologies used to address illegal third-party content and activity. In the early days of the commercial web, both the U.S. and the EU legal systems considered the idea of imposing general filtering mandates on intermediaries that required them to screen out illegal content, but ultimately rejected such mandates due to the significant burden on freedom of expression imposed by overbroad, imprecise filtering. Today, however, legislators are increasingly revisiting the presumption against filtering,[52] based in part on assumptions that filtering technologies have become more sophisticated. Proposed requirements for intermediaries to use filters are based on the rationale that, if there are suitable technologies available to address illegal content and the harms that may result from it, a refusal or unwillingness to use filters could be considered a form of negligence on the part of the intermediary. However, given the persistent risk to free expression posed by the automated detection and evaluation of speech, the deployment of filtering systems without regard for their harmful effects could also be considered a form of negligence. Intermediary liability laws should neither mandate, nor condition liability protection on, the use of filters.

Mandatory and voluntary use of automated systems

The mandatory use of particular forms of automation in content regulation raises prior restraint issues from a freedom of expression perspective. In fact, the mandatory application of such algorithmic systems, the automatic evaluation of content, and the subsequent restrictions on such content from being posted and disseminated fits the very definition of prior restraint, which is generally prohibited under freedom of expression doctrines. In view of this, lawmakers should refrain from imposing mandatory obligations on service providers to impose restrictions on speech through automated systems. Preference should be given to the voluntary deployment of such tools, in combination with safeguards for the fundamental rights of internet users, in view of the issues of false positives, due process, discrimination, and surveillance outlined above.

Systemic responsibilities for platforms and risk assessments

As noted in the paper on intermediary liability, some recent European laws and proposals move away from penalizing platforms for individual incorrect decisions about specific user expression, and instead seek to regulate platforms' overall content management operations and create new forms of administrative oversight with respect to these frameworks. The development of standards for transparency and accountability of content moderation practices is central to this.[53] One accountability mechanism that should be further explored is the use of human rights due diligence processes and other risk assessment methodologies. Risk assessment may be deployed with respect

to particular products or services that a platform develops as well as with particular known threats such as hate speech, terrorism content, and disinformation. Formal risk assessment procedures provide an opportunity for in-depth consideration of the potential impact on fundamental rights that a product or policy poses as well as the various measures that are and may be taken to address them.[54] It is crucial that freedom of expression and associated human rights, and the risks to the rights and freedoms of individuals, are included in these risk assessments. This prevents such risk assessments from being only focused on one particular set of harms and ignoring the potential harmful impacts of mitigating measures on freedom of expression and other fundamental rights.

Complaint and redress mechanisms

The use of automation in content moderation can give rise to additional problems of lack of due process for particular user content and activity. As a result, the development of new complaint procedures and new dispute resolution mechanisms has become a central part of the discussion about online content moderation and freedom of expression. Without proper complaint mechanisms, the actions that platforms take with respect to content and activity lacks accountability and endangers people's ability to effectively exercise their right to freedom of expression. The need for human oversight over automated decision-making deserves a particular emphasis in this context. It is paramount that the results of automated decision-making in individual cases can be scrutinized by humans and that wrong decisions are remedied, both at the individual decision level and through review of the systems that produced the error. Human oversight is necessary in all stages of the development and deployment of algorithmic systems. The possibility of human oversight when there are complaints about the function and impacts of algorithms for content moderation can provide a crucial safety net for the rights and freedoms of affected users.

Freedom of expression by design

"Regulation by design" has become an important way to safeguard fundamental rights. In the case of data privacy, regulation-by-design approaches have matured over the last two decades. These approaches have become an important foundation in privacy law and policy, underpinning data minimization strategies, information accountability, and privacy-friendly approaches to human computer interaction design. In the case of freedom of expression, regulation by design is still in its infancy. The development and deployment of AI and algorithmic systems more generally in the area of content moderation should draw lessons from the regulation-by-design literature developed in this area. By better incorporating freedom of expression and associated human rights concerns into the design and deployment of relevant tools and practices, the effective protection of freedom of expression can be better realized. To achieve this, relevant experts involved in the development and deployment of technologies for content moderation should be supported in developing a more robust field of "freedom of expression by design" approaches. In addition to transparency, accountability, and fairness, such approaches could extend to the creation of relevant datasets and labelling practices, as well as the particular ways in which algorithmic systems are being deployed.

Part II: Content Curation through Recommendation Algorithms

Introduction

Automation is used not only to moderate content, but also to recommend content to potential viewers. Important recommendation systems include content "feeds," such as Facebook's Newsfeed and YouTube's Recommended Videos, as well as search engines such as Google Search. These algorithmic tools play a central role in determining what content is seen online, and what remains hidden.

Content recommendations serve an important need in online media: helping people find relevant content. Based on certain personal and contextual information, these algorithmic systems search through an abundance of content to provide personalized selections of items that are (predicted to be) relevant for the user.

This is by no means a neutral process. Recommendation systems are designed and deployed by specific people in a specific context and for specific purposes. Accordingly, their content selections are dependent on a wide range of factors, including commercial and in some cases political considerations. A number of studies suggest that recommendation systems may funnel people toward disinformation, hate speech and violent extremism.[55] The following sections will elaborate on the functioning of recommendation systems, evaluate existing efforts to remedy their potentially radicalizing qualities, and offer alternative frameworks to mitigate their amplifying dynamics while minimizing the impact on freedom of expression.

Platforms also use (machine learning) algorithms to serve advertisements and other promoted content. This technology allows ad buyers, using large sets of personal data, to microtarget their advertisements toward highly specific audiences in ways that raise concerns about privacy, accountability, manipulation, discrimination, and bias.[56] Yet, despite both recommendation systems and advertising systems using algorithms to personalize their offerings, these types of content distribution are fundamentally different – in terms of substance, sources, functions, relevant harms, governance systems, applicable law, and so forth. Therefore, advertising and sponsored content exceeds the scope of this contribution, which focuses instead on content recommendations for *organic* content, i.e., content that is disseminated without payment to the platform.

Explaining recommendation systems and their deployment

Recommendation systems are automated tools that present ("curate") a selection of content ("recommendations") from an abundance of content. These personalized selections are the result of two main processes. The first is the collection of information about an individual user or browser and the development of a profile based on that information. This involves analyzing data from the user and others like them (website visits, articles read, social media behavior such as clicks and likes). This data is sometimes received from third-party data brokers who sell the data to the entity creating the profile. The second process involves matching the user with the larger pool of content from which the recommendations are drawn. The software system computes a similarity score between the user profile and the characteristics of each individual content item.[57] Recommendation systems are presented as tools to help users find more "relevant" content, but they can also serve other goals.

Most importantly, they help content providers and platforms generate advertising revenue by increasing user engagement.[58]

Proprietary concerns related to this process may make it hard to pinpoint what recommendation systems are actually composed of, and how they function. In turn, this creates serious complications and limits to solutions aimed at transparency.[59] Platforms may be unwilling to lay bare the workings of their recommendation algorithms for proprietary reasons. Even without this impediment recommendation systems are complex systems to grasp, explain, and hold accountable,[60] and some have argued that this narrative can also be a deflection strategy employed by platforms to avoid scrutiny of their systems.[61]

The way recommendation systems work differs between platforms. While search engines like Google deliver filtered selections in response to user queries, social media platforms like Facebook and Twitter give personalized feeds of content independent from any explicit user input. These active and passive recommendation systems both filter out and prioritize content according to specific algorithmic procedures that are optimized for personal "relevance." However, what relevance *means* is not always explained in detail to the public, allegedly because these algorithmic relevancy formulas are well-kept corporate secrets.[62] Users might get *some* explanation as to why they receive a certain post, but the level of detail varies between platforms and services. In some cases, relatively detailed information is published and can offer helpful guidance for sophisticated users to optimize their content for findability.[63] But other services operate largely as "black boxes," certainly when it comes to explaining these complex systems in a meaningful way to the average internet user. Basic disclaimers and notices ("you see this post because you like politics") may not be not very informative about the precise selection criteria at work. What their secret recipe involves, precisely, is often unclear. As discussed further below, a variety of interests and considerations play a role here: recommendation systems are designed for audience engagement, but increasingly also as a vehicle for content moderation or curation.[64]

Recommendation systems are widely used today in various contexts, ranging from commerce (Amazon), travel (Booking.com), music (Spotify), and video on demand (Netflix) to, most relevant for a discussion of disinformation and hate speech, social media platforms (Facebook, Twitter, and YouTube). These social media platforms rely on advertising revenue and their business model centers on engagement, or keeping users "hooked." Controversial, provocative, and extreme content can drive engagement, so their algorithms learn to prioritize such content when it is popular with users.[65]

It should be noted that recommendation systems do not work in isolation, but interact with other human and non-human actors in complex ways. First, platform recommendation systems depend on the content uploaded and shared by users, including content which may be contentious, harmful, or unlawful. The pool of content from which recommendation systems draw is not a neutral representation of public opinion, but biased toward the interests of its most active contributors.[66] And this content pool can tend toward fringe views, if only because there is a relatively weaker incentive to share content displaying scientific or societal consensus.[67] For example: one will find less content arguing that the earth is round, that vaccinations are safe, or that the Clintons are not pedophiles or that there is no relation between race and IQ. The opposing views, however, are overrepresented. This skews recommendations toward the extreme content that is available, even

with hypothetically neutral algorithms. Indeed, research shows that sophisticated users can exploit "data voids" in the content supply: by targeting obscure terms and topics with little to no existing results, they can attempt to draw attention toward potentially manipulative or harmful content.[68] Second, users also influence recommendation systems by behaviors such as liking, rating, sharing, following, scrolling, reading, and clicking; their algorithms are tuned to take popularity (or virality) into account.[69] Strategic actors can therefore deploy bots which send, (re)tweet, like, and spread contentious content to artificially amplify the popularity of certain topics.[70] In addition, research has shown that "false news" spreads much faster and further than "truthful" news, at least within certain user populations.[71]

All this means that there is no such thing as a completely neutral, objective, or unbiased recommendation system: they necessarily reflect and amplify certain preferences, biases, and intentional distortions introduced by their users. In this regard, content recommendations are not entirely unlike editorial decisions made in traditional media; they involve a judgement of relevance and newsworthiness that is necessarily value-laden.[72] (Of course, editors and recommendation systems also differ in many regards, including that recommendation systems are automated and process third-party content, and as a result are generally less intentional or deliberate about overall outcomes.)

The effects of recommended content are highly unpredictable. Nevertheless, actors with more oversight, resources, and/or knowledge of platform dynamics are generally better able to manipulate the wider information landscape. This threat became visible during the "fake news" controversy around the 2016 U.S. presidential election and UK's Brexit referendum, and the subsequent Cambridge Analytica/Facebook scandal,[73] after which it became clear that certain governmental actors had strategically exploited the underlying dynamics of platforms and intentionally inserted dubious claims and outright propaganda in the media ecosystem for political and/or commercial goals.[74] The effects of such interventions on public opinion and politics are difficult to measure or quantify. But these stories do underscore that amplification of harmful content through recommendation systems is not necessarily fully intentional or expected on the part of the platform; it may occur without the platform's full knowledge, intent and control, since these systems operate in complex and dynamic networks of multiple invisible actors and incentives.[75] Of course, deploying these systems without proper research into their potentially harmful effects could still raise a charge of negligence or recklessness.

The complex role of recommendation systems in online hate speech and disinformation

Research on the role of recommendation systems in the circulation of disinformation and hate speech online has surged in recent years. However, studying these systems has proven difficult because of the complexity and magnitude of the current information ecosystem and its ever-changing recommendation algorithms, and the limited cooperation from social media platforms with research communities. Despite these challenges, more researchers within academia and beyond are developing ways to measure and understand how recommendation algorithms contribute to the spread of (dis)information.

Wittingly or not, platforms may actively contribute to the amplification of incendiary, controversial and divisive (dis)information as it directly aligns with the commercial and technological

infrastructures of their recommendation systems that are optimized for user engagement. However, blaming recommendation systems alone ignores the fact that these infrastructures work in conjunction with users' own biased content and behavior, and are furthermore used and strategically exploited by sophisticated actors with more resources and experience than the average user, who can accordingly work the system and gain more political influence.

To give a few examples: Guillaume Chaslot, a former YouTube engineer, developed an algorithmic method to show which recommendations YouTube gives on certain popular topics.[76] He showed that by inserting rather ordinary queries, people increasingly get more and more extreme items recommended, in part due to the content-availability asymmetries noted earlier: asking about "vaccine facts," it takes only a few steps to get to anti-vaxxer conspiracy theories; "global warming" to climate change denialism; "US presidential election" to pro-Trump videos, and entering "the pope" gives suggested videos describing the Catholic leader as "evil," "satanic," or "the anti-Christ."[77] Chaslot's samples are not representative, but nonetheless provide a revealing snapshot that lends empirical support to the theory that YouTube is a "great radicalizer."[78] Since these events, YouTube has revised its algorithms in an attempt to counter extremism.[79]

These findings have been corroborated by a number of other, non-U.S.-focused studies. Kaiser and Rauchfleisch (2017) report on how people watching videos of the populist right-wing party Alternative für Deutschland are recommended by YouTube to watch videos by the vastly more extreme and openly anti-Semitic National Democratic Party of Germany (NPD).[80] Social-media analyst Ray Serrato used computational methods to study the recommendations given by YouTube when searching for "Chemnitz," the East German city where violent anti-immigrant protests erupted in 2018.[81] He shows how ordinary viewers searching for this term were led by YouTube toward more and more extreme videos, while a tightly networked ecology of users and channels was able to amplify the reach and virality of right-wing videos. And investigative journalists from the Dutch news outlets de Correspondent and de Volkskrant undertook a major study into YouTube's "radicalization problem."[82] They gathered massive amounts of data (660,000 videos from 1,500 channels, with 120 million reactions, 15 million recommendations, and 440,000 video transcripts). Their analyses revealed the presence of a tightly knit right-wing reactionary network on YouTube, which lured viewers into a right-wing maze of more and more extreme videos with the help of YouTube's recommendation algorithms. They also showed the radicalization process of certain "heavy commenters" based on content analyses of their comments over a yearlong period. Their efforts to study the radicalizing effects of the recommendation algorithms did not yield conclusive results, which they refrained from publishing.

These studies reveal the difficulty of isolating the influence of recommendation algorithms in the amplification of hate speech and disinformation from the strategic actions of those who use YouTube's platform to convince viewers of their ideology. This intricate entanglement of sociological and technological factors is particularly made clear by the research done by Rebecca Lewis at the Data & Society Research Institute (2018). She identifies a wide, interconnected network of "alternative political influencers" on YouTube who promote alt-right ideologies with specific online branding techniques (testimonials, controversy, platform optimization). Having analyzed 65 influencers across 81 channels, she argues that such influencers facilitate radicalization through social networking: by referencing others in the network, and through guest appearances on each other's

shows, they let audiences move from mainstream conservative to more and more extreme right-wing contents. She explains their appeal by describing how they establish an alternative sense of credibility based on personal authenticity and relatability, and cultivating a social identity of a countercultural underdog.

Over the past two years, social media platforms have started adjusting their recommender algorithms in a bid to suppress harmful content. A key question going forward is whether these changes will have their intended effect, and what other (unintended) consequences they might have for online discourse.

Analysis: How are platforms and governments addressing the algorithmic amplification of hate speech and disinformation?

Recommendation systems are currently subject to a range of efforts to combat the amplification of disinformation and hate speech. These include (proposed) government interventions, as well as self-regulatory initiatives among platforms and/or civil society. Generally speaking, their toolbox includes the following interventions: content removal and other forms of moderation, algorithmic content curation, user customization options, transparency, and media literacy.

Content removal, demonetization, and/or downranking

One way to respond to amplification concerns is to remove the content at issue. Of course, such content moderation is by no means straightforward. Content removal is not always effective at combating the perceived harms or identifying the content at issue, and raises questions about gatekeeping and the freedom of expression. When it comes to amplification, content removal is especially limited because it may concern content that does not necessarily violate content policies. For instance, one might be concerned about the lack of political balance on a given recommendation system without wishing to actually prohibit content from a particular political orientation or origin. Removal, in other words, is a relatively extreme measure and may not always be proportional.

Platform moderation consists of removing dubious actors and content. In theory, this can help to stop the spread of disinformation, hate speech, and violent extremism. In practice, this is a complex exercise: how are these takedown decisions implemented, following what rules and procedures? Can the speakers object or intervene, and if so how? How are human rights (freedom of speech, rights to information) guaranteed when private actors (platforms) have such censoring powers? How are critical voices not marginalized? Are platforms in any way accountable for their moderation practices, and how can or should their decision-making be publicly scrutinized? Research shows that content takedown is not always effective, either: much of the harmful content and actors remain active, giving them a false aura of legitimacy.[83]

Content moderators have other tools than removal. For instance, YouTube demonetizes some content, terminating or suspending any revenue sharing agreements with the content provider. This can be a powerful deterrent, since many YouTubers rely on ad revenue as a key source of income, and raises comparable concerns from a free speech perspective, at least for professional content producers. By cutting off a key source of revenue, platforms can render certain content unviable to produce, and effectively silence certain speakers.

Another alternative to content removal is downranking: the harmful content is deprioritized in news feeds and other recommendation systems, so that it becomes less visible and less likely to be amplified. This is where the content moderation debate intersects with the broader issue of algorithmic content curation: what design principles are applied in recommendation systems, and how does it pick winners and losers? This is discussed in detail below.

Algorithmic content curation (and non-discrimination)

As discussed, recommendation systems are not inherently neutral and are designed to prioritize content with certain characteristics and deprioritize other content. As such, their functioning is already a form of content moderation: by suggesting some types of content and hiding others, they perform an important gatekeeping function. A number of governments are now proposing to have platform recommendation algorithms accommodate public interest considerations and legal requirements, and platforms are taking comparable measures on their own initiative. For instance, the European Commission's Code of Practice on Disinformation requires platforms to "[d]ilute the visibility of disinformation by improving the findability of trustworthy content" and to "invest in technological means to prioritize relevant, authentic, and authoritative information."[84] The Council of Europe emphasizes the importance of diversity. Its Committee of Ministers has called on member states to foster partnerships between social media platforms and outside stakeholders "to enhance users' effective exposure to the broadest possible diversity of media content."[85] Many of these "public interest considerations," however, are relatively undertheorized and underdeveloped, in terms of providing guidance that can be operationalized and evaluated in algorithmic systems.

Platforms are taking comparable measures of their own accord. In 2018 alone, Facebook announced updates to promote content from friends and reduce the reach of news pages; to downrank "false news" content flagged by accredited fact-checkers; and to downrank "borderline content" that falls short of violating company policies.[86] In April 2019, Facebook started downranking anti-vaccination content.[87] In May 2019, Facebook presented its new "click-gap" method to suppress "low-quality content."[88] Similarly, in January 2019, YouTube announced that it would "begin reducing recommendations of borderline content and content that could misinform users in harmful ways."[89]

Another set of efforts focuses instead on non-discrimination rules, which would place limits on such algorithmic curation. For instance, the German federal broadcasting authority has proposed, in an instrument known as the *Medienstaatsvertrag,* to prohibit platforms from discriminating against "journalistic editorial content" to the extent that the intermediary has "potentially a significant influence on their visibility."[90] In the Netherlands, the Dutch State Commission on the Parliamentary System proposed a comparable "independent entity" to monitor platform recommendations, but unlike the Germans, its mandate would not focus on non-discrimination but rather on maintaining "diversity" and avoiding "bias."[91] In the U.S., Senator Josh Hawley (R-MO) has proposed that platforms observe "political balance" in their algorithms.[92] However, implementing such principles in practice is not straightforward: as discussed, it is not clear how recommendation algorithms can be made "neutral," or what would constitute "discrimination," since they necessarily rank some content over others. In some sense, the entire purpose of these algorithms is to discriminate. Even more difficult are concepts like "political balance," "bias," and "diversity," which implicate value-laden and content-specific judgements about newsworthiness. Indeed, these two types of principles

(non-discrimination and diversity) may contradict each other; diversity rules could require platforms to prioritize certain specific types of content, whereas non-discrimination rules might prohibit them from doing so.

Without further elaboration, therefore, these regulatory standards seem immature at best, and implementing them would be both technically infeasible and politically controversial. Particularly with such vague and subjective rules, government action in this space also raises questions about freedom of expression and the rule of law. Without adequate safeguards, government regulation of content recommendation could impede the freedom of expression of social media users. Prescribing what should be downranked risks becoming a form of censorship, and what must be prioritized a form of propaganda.

Finally, it is worth noting that antitrust and consumer protection law already place some limits on algorithmic curation and discrimination. First, antitrust law could place limits on the ability of platforms to prioritize their own services (as seen in the EU case against Google's Search and Shopping products). Secondly, consumer protection law in most jurisdictions forbids covert advertising, which means that platforms have a duty to disclose whether content is being sponsored – i.e., whether it constitutes an advertisement rather than organic content. Such disclosures are already common practice on most major platforms, but they are not as visible and recognizable as they could be: advertisements are often incorporated into organic content feeds, and their design risks blurring the boundaries between these two types of content.

As noted, platforms also respond to public pressure and changed their recommendation algorithms to prioritize "trusted sources" (whitelisting) and deprioritize "harmful content" (blacklisting), which runs into the same problems as content moderation.[93] Downranking also raises many of the same free speech issues as content moderation: it prevents platform users from effectively making their voices heard, with little to no accountability when their content is removed. Because of "proprietary reasons," platforms are not transparent about what such changes in their recommendation algorithms look like, hindering public scrutiny and accountability. Some argue that as long as platforms work on ad-based models, they have an incentive to permit, or at least minimize, their investments in combating disinformation and other incendiary content because this content keeps people engaged on their platforms and thereby creates a source of revenue.[94] Public pressure may still motivate platforms to act (or be seen to act), but these countervailing incentives should be taken into account when assessing the prospects of self-regulation in this space.

User customization options

Another policy option is to develop tools for user choice, so they can customize their recommendation systems. The aforementioned Council of Europe recommendation calls on states to encourage platforms to "provide clear information to users on how to find, access and derive maximum benefit from the wide range of content that is available."[95] Similar rules are found in the European Commission's Code of Practice and the German broadcaster's proposals.

In practice, platforms already provide several options for user customization. Basic functionalities such as liking, following, and subscribing can be seen as a form of user choice, since they help users to express what kind of content they would like to receive. YouTube and Facebook allow users to

express *disinterest* and filter out certain kinds of content or sources. Twitter allows users to view tweets in a chronological order to avoid additional algorithmic curation. However, users may not always be aware of or interested in these features – particularly when they are not easily visible and accessible in the platform's visual interface.

Transparency

Perhaps the most widely supported policy priority around recommendation systems is transparency. All of the aforementioned policy instruments include transparency obligations of some sort. Further transparency obligations can also be found in horizontal instruments. The General Data Protection Regulation grants data subjects the right to demand "explanations" about recommendation systems.[96] Similarly, the EU's Regulation on Promoting Fairness and Transparency for Business Users of Online Intermediation Services (Platform-to-Business Regulation) requires platforms to explain "the characteristics of the goods and services offered to consumers through the online intermediation services or the online search engine." The revised Audiovisual Media Services Directive has additional rules specifically for video platforms. However, transparency is a broad term and can take many forms -- particularly in the context of technically complex systems like content recommendations. These issues of transparency are given particular attention in a separate Transatlantic Working Group document.[97]

Media literacy

Last but not least, interventions can target users themselves: improving media literacy. Most, if not all, European states have programs, often with government support, aimed at increasing people's ability to discern information quality. Such programs can include practical skills, critical reasoning, and ethical considerations. The idea is simple: education better enables people to navigate today's complex media ecosystems and to be more resilient against the propaganda campaigns of malicious actors. Media literacy is recognized as a fundamental skill in the 2016 Audiovisual Media Services Directive, a key document defining the EU's future media and communications policy. The European Commission supports initiatives and prizes, manages projects, programs, and funding schemes, such as Creative Europe, coordinates with member states on policies and best practices, and develops new policies based on expert group findings. Each individual member state has similar programs that tackle this issue from country-specific perspectives, often in conjunction with civil society and educational initiatives.

Raising awareness and media literacy is useful and important; because these automated systems are both novel and complex, they are poorly understood by many users. Helping them to understand how recommendations are generated and what they can do to alter and customize their experience can help users to act with greater care and autonomy in the face of harmful or misleading content offerings. Still, this approach is limited by the fact that some people willfully engage in spreading contentious content. Sociological research shows that these people are attracted by disinformation not necessarily because they consider it truthful, but rather because it aligns with their worldview and it gives them a sense of community and identity.[98] A related problem is that those in most need of "education" may not be the ones who are actually reached and most receptive to it. Therefore, educating citizens may not always work in combating disinformation in society at large because the issue is not just cognitive – i.e., based on a faulty understanding of social media or recommender

algorithms – but is also driven by deeper and more complex cultural and societal shifts related to the loss of trust in mainstream media, science, and other knowledge institutions.

Fact-checking

Tracing, highlighting, and correcting disinformation is important for reasons of transparency and truth-finding, but is not a comprehensive solution either. At a practical level, fact-checking is difficult to perform at scale: determining truth is one of the most difficult aspects of content moderation, even for trained fact-checkers.[99] Even if possible, it will be properly difficult to automate and it will continue to require human review, which is costly and time-consuming.[100] More fundamentally, as discussed, the truthfulness of disinformation is often of secondary importance to those producing and sharing it: it is mainly about worldview and identity, not truth. Accordingly, fact-checking corrections will not always be taken seriously by people who ideologically do not align with its findings.[101] Moreover, because fact-checkers are often from opposed ideological and societal groups, these organizations may suffer a lack of trust among their target audience.[102] Indeed, highlighting false information may even be counterproductive, since the excessive attention (and uptake by other media organizations) can also increase its reach.[103] It has even been argued that being confronted with corrections could actually further *strengthen* the original beliefs.[104]

(Self) Regulation

Various regulatory frameworks have been discussed to tackle the amplification of disinformation and hate speech, ranging from aforementioned efforts to impose more diverse recommendations to stricter rules for content moderation.[105] However, governments have been reluctant to take a strong lead here. For now, (supra)national governments have tried some forms of co-regulation, such as the 2018 European Commission's Code of Practice, in which the major tech companies pledged to work more actively to lessen the spread of disinformation and hate speech online. However, all of this is non-binding and the rules of the game can be rather freely interpreted by these companies. Various critics in academia and civil society argue that such forms of self-regulation do not provide enough incentives for platforms to make meaningful changes, which might threaten their core business models. Fact-checking organizations and academic researchers who partnered with Facebook have expressed dissatisfaction at the lack of progress in these arrangements. Some argue therefore to regulate recommendation systems themselves.[106] The development of effective regulations may be challenging, not only due to its technical complexity and dynamism but also due to its extreme political sensitivity, and its implications for fundamental rights including freedom of expression.

Recommendations

Despite these difficulties in preventing the amplification of disinformation and hate speech online, it remains important to think about other possible remedies and solutions. We propose here some potential directions and evaluate their feasibility.

Raising awareness and transparency

While transparency of recommendation systems is surely no panacea, and it knows many pitfalls,[107] there is much to gain by raising awareness of their functioning. Transparency can help to hold these systems accountable and enable more evidence-based policy making. To this end, governments could

impose stricter enforcement of users' right to explanation under the GDPR, and require platforms to offer additional forms of transparency of their recommendation systems. This could take several forms, including user-facing notices, government or civil society auditing, academic partnerships, and regimes for public disclosure (discussed in the TWG's working paper on transparency).

One particularly important form of transparency is the sharing of data and information with outside researchers. First, non-sensitive, anonymized data could be shared in public datasets.[108] Second, sensitive data can be shared in partnerships with relevant institutions, under non-disclosure disagreements to safeguard confidentiality. Given the technical and legal difficulties faced by self-regulatory initiatives in this space, such as Social Science One, there may be a role for governments to facilitate these exchanges (e.g., by providing a processing ground under data protection law, and/or by imposing sanctions for breaches of confidentiality).

To improve awareness, legacy media and civil society organizations should pay more attention to the social and cultural contexts in which people radicalize, rather than just criticizing social media platforms. Reporting on recommendation systems should not lose sight of the bigger picture. For instance, the Mozilla Foundation took up an article by the Washingtonian, which reported on a mother and her teenage son who turned to white supremacist movements online after being falsely accused of a sex offense at high school.[109] Mozilla brought the mother and son together and invited their community to pose questions to them. Personal stories like these help our understanding about radicalization processes beyond a sole focus on the radicalizing powers of recommendation systems.

Increasing user control

Some scholars argue that increasing user control in recommendation systems would mitigate many of the aforementioned problems, and enhance the individual and societal value of recommendation systems.[110] Besides empowering users to make recommendation systems more responsive to *their* interests and needs, and not what platforms think is most relevant to them, user control simultaneously requires recommendation systems to be more transparent and explainable. Not only can this improve user satisfaction and trust, it can also counteract "filter bubble" concerns by encouraging them to look beyond their assumed or known interests. Users could, for example, be confronted with explicit options to receive recommendations outside their ordinary consumption habits, like a "get me out of my filter bubble button" or a "show me more of the ideological other side" slider. There are several ways to increase user control, at the input level of user preference and in choosing from different kinds of recommendation algorithms so as to get different *kinds* of recommendations depending on one's mood or information interests at a given moment.[111] This being said, some would warn that increasing user control can also serve to enable users who deliberately choose to view extremist or contentious content. Much depends on how user control is implemented and designed, and more empirical research (and access to data) is needed to study the effects of these tools in practice.

Multistakeholder governance

In designing and implementing these interventions, one way to mitigate free speech concerns regarding social media regulation is to incorporate multistakeholder and co-regulatory elements.[112] For instance, the Council of Europe's guidelines related to media pluralism policy recommend that

social media platforms and other online services engage in "open, independent, transparent, and participatory initiatives" alongside "media actors, regulatory authorities, civil society, academia and other relevant stakeholders" on such issues as algorithmic diversity and transparency.[113] Such initiatives could assist in many of the efforts discussed above, including the design of recommender systems and algorithmic curation as well as interventions for transparency and media literacy.

Bringing different stakeholders together, instead of relying exclusively on the judgment of either dominant platforms or government regulators, can create checks and balances on these powerful actors while helping them take account of outside viewpoints, interests, and expertise. In this way, opening up the governance process to experts and affected stakeholders has the potential to enhance both its effectiveness and its legitimacy.[114]

Designing effective and inclusive regulatory institutions of participation is a key challenge going forward; on the one hand, strictly voluntary arrangements may fail to hold powerful commercial actors accountable. Government, therefore, may have an important role to play in undergirding these initiatives with the force of law. On the other hand, the participation of outside experts should not serve as a mere fig leaf on government policy; to have its intended effect, multistakeholder governance should meaningfully involve other actors in the decision-making process. These challenges may be serious, but inspiration can be drawn from earlier precedents in media governance; for instance, lessons can be drawn from the design of self-regulatory and co-regulatory bodies in journalism, advertising, and public broadcasting. Another relevant avenue of debate is the concept of "social media councils," explored in a separate Working Group report.[115]

Incentives for alternative business models

As discussed, ad-based business models give social media platforms a strong commercial incentive to keep users engaged, including via controversial and incendiary contents. Some critics have therefore argued to push or incentivize platforms toward a subscription-based model.[116] Subscription-based models have the advantage of a revenue stream irrespective of advertising, meaning that recommendation systems need not be designed to optimize for engagement only. Netflix and Spotify, the two largest subscription-based platforms, increasingly use other metrics, such as user satisfaction and explorability as well. Although such an approach introduces social equity issues (people with lower incomes may not be able or willing to pay for ad-free services), it could thus make good sense to think of financial or regulatory incentives to push ad-based platforms to other business models with fewer incentives to push inflammatory content. One starting point, for instance, might be to impose organizational restrictions to reduce the influence of advertising operations' on organic content recommendations. Lessons could be drawn from best practices in journalism, where editorial decision-making is also designed to be independent from commercial operations. These are evidently more far-reaching interventions than the aforementioned, and have not yet received as detailed a treatment in policy research and debate. Further research would therefore be needed to explore the potential applications and viability of these more muscular approaches to social media regulation.

Conclusion

Platforms use countless different forms of automation to shape our experiences online. As we have seen, this is not just about "AI." And although technological solutions for content moderation and curation are increasingly widespread, many are still rudimentary and imperfect. Use of automation in content moderation exposes all speech to a form of evaluation *ex ante* and in a way that fails to consider linguistic, social, historical, and other relevant context – which creates substantial risks to the freedom of expression. Governments should therefore act with caution and resist simplistic narratives about all-powerful algorithms or AI as being the sole cause of, or solution to, the spread of harmful content online. Indeed, any legal requirements to adopt specific forms of automation are likely to be premature, and would present major risks to freedom of expression. For now, what governments should focus on is enhancing transparency in existing practices, empowering research communities with the necessary data, and ensuring that users have access to meaningful choice and redress mechanisms.

Notes

[1] Emma Llansó is director of the Center for Democracy & Technology's Free Expression Project, which works to promote law and policy that support internet users' free expression rights in the United States and around the world.

[2] Joris van Hoboken is a senior researcher at the Institute for Information Law (IViR), and a professor of law at the Vrije Universiteit Brussels (VUB). He works on the intersection of fundamental rights protection (data privacy, freedom of expression, non-discrimination) and the governance of platforms and internet-based services.

[3] Paddy Leerssen is a Ph.D. candidate at the Institute for Information Law (IViR). He focuses on the European governance of recommendation algorithms in social media platforms and their impact on media pluralism – how platforms can be made publicly accountable

[4] Jaron Harambam is an interdisciplinary sociologist working on news, disinformation, and conspiracy theories in today's algorithmically structured media ecosystem. He currently holds a Marie Sklodowska-Curie Individual Fellowship at KU Leuven's Institute for Media Studies.

[5] For example, one set of estimates from 2018 found that, every minute, users post nearly 2,000 comments on Reddit, 50,000 photos on Instagram, 80,000 posts on Tumblr, 470,000 tweets on Twitter, 2 million snaps on Snapchat, and 3.8 million search queries on Google, Domo, Data Never Sleeps 6.0, https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf.

[6] A note on terminology: The phrase "artificial intelligence" broadly refers to computer systems that can perform tasks associated with intelligent beings. Its use in policy-making processes is typically imprecise, evoking powerful technical capabilities without necessarily specifying existing technical processes. "Machine learning" is a branch of computer science focused on computer programs that adapt to and "learn" how to act from data without being specifically programmed by a human.

[7] Use of AI in Content Moderation, Cambridge Consultants for UK OfCom (2019) at p. 37, https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

[8] *See* Ekram Sabir et al., Recurrent Convolutional Strategies for Face Manipulation Detection in Videos, https://arxiv.org/abs/1905.00582.

[9] This use of unsupervised learning to create labeled inputs for a supervised-learning process is sometimes called "self-supervised" learning. *See* Andriy Burkov, https://www.kdnuggets.com/2019/01/burkov-self-supervised-learning-word-embeddings.html.

[10] E.g. T. Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (July 2016), https://arxiv.org/abs/1607.06520.

[11] https://towardsdatascience.com/the-a-z-of-ai-and-machine-learning-comprehensive-glossary-fb6f0dd8230

[12] https://developers.google.com/machine-learning/glossary#recurrent_neural_network

[13] Mixed Messages p. 9, https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

[14] "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," Felbo et al., https://arxiv.org/pdf/1708.00524.pdf

[15] https://www.perspectiveapi.com/#/home

[16] https://github.com/conversationai/perspectiveapi/wiki/perspective-hacks

[17] https://docs.coralproject.net/talk/toxic-comments/

[18] Deceiving Google's Perspective API Built for Detecting Toxic Comments, Hosseini et al. (2017) https://arxiv.org/pdf/1702.08138.pdf.

[19] The Risk of Racial Bias in Hate Speech Detection, Sap et al. (2019), https://www.scribd.com/document/421898931/The-Risk-of-Racial-Bias-in-Hate-Speech-Detection

[20] https://conversationai.github.io/

[21] https://openai.com/blog/better-language-models/. Specifically, "In order to preserve document quality, we used only pages which have been curated/filtered by humans—specifically, we used outbound links from Reddit which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting (whether educational or funny), leading to higher data quality than other similar datasets, such as CommonCrawl."

[22] Language Models are Unsupervised Multitask Learners, Radford et al. (2019) https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. See also https://nostalgebraist.tumblr.com/post/187579086034/it-seems-pretty-clear-to-me-by-now-that-gpt-2-is

[23] https://openai.com/blog/better-language-models/

[24] Release Strategies and the Social Impacts of Language Models, Solaiman et al. (2019) https://arxiv.org/ftp/arxiv/papers/1908/1908.09203.pdf. *See also*, OpenAI has released the largest version yet of its fake-news-spewing AI, Karen Hao, (Aug. 29, 2019), https://www.technologyreview.com/s/614237/openai-released-its-fake-news-ai-gpt-2/

[25] OpenGPT-2: We Replicated GPT-2 Because You Can Too, Aaron Gokaslan and Vanya Cohen et al., (Aug. 26, 2019) https://blog.usejournal.com/opengpt-2-we-replicated-gpt-2-because-you-can-too-45e34e6d36dc

[26] If the goal is to ensure the integrity of the message, as in cryptographic hashing, then this sensitivity of the hash to the most minute change is the goal of running the hash function. Perceptual image hashes are much more vulnerable to attacks that can reveal information about the content that has been hashed. *See* https://towardsdatascience.com/black-box-attacks-on-perceptual-image-hashes-with-gans-cc1be11f277

[27] *See* Perceptual video hashing based on the Achlioptas's random projections, R. Sandeep and Prabin K. Bora, https://ieeexplore.ieee.org/document/6776252

[28] https://jenssegers.com/perceptual-image-hashes

[29] https://web.archive.org/web/20130921055218/http://www.microsoft.com/global/en-us/news/publishingimages/ImageGallery/Images/Infographics/PhotoDNA/flowchart_photodna_Web.jpg

[30] https://medium.com/@timanglade/how-hbos-silicon-valley-built-not-hotdog-with-mobile-tensorflow-keras-react-native-ef03260747f3

[31] https://gizmodo.com/british-cops-want-to-use-ai-to-spot-porn-but-it-keeps-m-1821384511

[32] Note the difference between face detection, which identifies the presence of (typically human) faces in an image, and facial recognition, which attempts to match a detected face to an identified person.

[33] Louise Matsakis, Tumblrs Porn-Detecting AI Has One Job--And It's Bad At It, https://www.wired.com/story/tumblr-porn-ai-adult-content/ (Dec. 5, 2018).

[34] OfCom 2019, 51.

[35] For example, Facebook has announced plans to develop a dataset for researchers to use in developing technology to detect deepfakes. Facebook, "Creating a dataset and a challenge for deepfakes" (Sept. 5, 2019) https://ai.facebook.com/blog/deepfake-detection-challenge/

[36] *See* Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, Explaining and Harnessing Adversarial Examples (2015) https://arxiv.org/pdf/1412.6572v3.pdf (example of image transformation that leads classifier to conclude image of panda is a gibbon).

[37] Nicholas Papernot et al., Practical Black-Box Attacks against Deep Learning Systems Using Adversarial Examples (2016), https://arxiv.org/pdf/1602.02697v2.pdf

[38] https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed

[39] See Adrian Yijie Xu, AI, Truth, and Society: Deepfakes at the front of the Technological Cold War, (July 2, 2019) https://medium.com/gradientcrescent/ai-truth-and-society-deepfakes-at-the-front-of-the-technological-cold-war-86c3b5103ce6

[40] Id.

[41] https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography

[42] Joe Uchill, Why the deepfakes threat is shallow (Aug. 15, 2019), https://www.axios.com/why-the-deepfakes-threat-is-shallow-16caf6a0-af83-4dbc-9008-6a2d4a2f08ae.html

[43] *See,* Camille François, Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses, https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf

[44] *See* Mixed Messages, *supra* n.13, p.14-15.

[45] http://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/

[46] OfCom 2019, p. 26.

[47] See, e.g., Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx

[48] See, e.g., the recent Declaration (Feb. 13, 2019)1 on the manipulative capabilities of algorithmic processes. The Council of Europe is finalizing a round of consultations on its draft recommendation on the human rights impacts

[49] Ananny, M. (2019). Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance. *Free Speech Futures, An essay series reimagining the First Amendment in the digital age*, Columbia University. ("Today, the meaning and force of the First Amendment play out in the new and often unstable technological infrastructures and institutional spaces of social media platforms.")

[50] Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/ReportGA73.aspx

[51] *See* Joris van Hoboken and Daphne Keller, Design Principles for Intermediary Liability Laws, TWG Discussion paper, October 2019, available at https://www.ivir.nl/publicaties/download/Intermediary_liability_Oct_2019.pdf.

[52] *See*, for instance, Article 17 of the EU's new 'Copyright Directive': Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. OJ L 130, available at: http://data.europa.eu/eli/dir/2019/790/oj.

[53] *See, for example*, Mark McCarthy, Transatlantic Working Group paper

[54] *For example*, Google conducted, and published, a human rights impact assessment for its Celebrity Recognition API (October 2019): https://services.google.com/fh/files/blogs/bsr-google-cr-api-hria-executive-summary.pdf. Facebook has since published a human rights impact assessment of its newly formed Oversight Board: https://bsr.app.box.com/s/8r0vw4a5kib6y6xfdtd5jin3g3fcbzs5.

[55] *See* Part II Explaining recommendation systems and their deployment below (Under "The role of recommendation systems in online hate speech and disinformation").

[56] Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *Arxiv 1904.02095v4 [Cs]*. Retrieved from https://arxiv.org/pdf/1904.02095.pdf

[57] Ricci, F., Rokach L., & Shapira, B. (2015) *Recommender Systems Handbook*. Springer, Cham, SH.

[58] Ibid.

[59] Ananny, M., & Crawford, K. (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3): 973-989.

[60] Diakopoulos, N. (2015) Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3): 398-415.

[61] Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133).

[62] Burrell, J. (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3: 1.; Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

[63] For instance, Google provides relatively detailed guidance to 3rd party reviewers that evaluate search results:: https://www.google.com/search/howsearchworks/mission/users/ and https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

[64] *See* Part II Analysis: How are platforms and governments addressing the algorithmic amplification of hate speech and disinformation? below (Under 'Algorithmic content curation (and non-discrimination').

[65] Cobbe, J., & Singh, J. (2019). Regulating Recommending: Motivations, Considerations, and Principles. (April 15, 2019). Available at SSRN: https://ssrn.com/abstract=3371830 or http://dx.doi.org/10.2139/ssrn.3371830; Gary, J., and Soltani, A. (2019) *First Things First: Online Advertising Practices and Their Effects on Platform Speech*, Knight First Amendment Institute at Columbia University, available at: https://knightcolumbia.org/content/first-things-first-online-advertising-practices-and-their-effects-on-platform-speech; Lewis, P. (2018). Fiction is outperforming reality": How YouTube's algorithm distorts truth. *The Guardian*, February 2, 2018. A recent study on YouTube's algorithms refutes the radicalization claim and finds evidence that YouTube's recommendation algorithm favors mainstream sources. See Ledwich, M., & Zaitsev, A. (2019). Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization. *arXiv preprint arXiv:1912.11211*.

[66] *See* M Golebiewski and D Boyd, Data Voids: Where Missing Data Can Easily Be Exploited (2018), https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

[67] Bahara, H., Kranenberg, A., Tokmetzis, D. (2019) Hoe YouTube rechtse radicalisering in de hand werkt, *De Volkskrant*, 8 februari 2019.

[68] *See* M. Golebiewski and D. Boyd, Data Voids: Where Missing Data Can Easily Be Exploited (2018), https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

[69] Napoli, P. (2014). Digital intermediaries and the public interest standard in algorithm governance. *Media Policy Blog*.

[70] Bennet and Livingston, 2018

[71] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353-369.

[72] *See* Helberger, N. (2019, in press). On the democratic role of news recommenders, *Digital Journalism*. Available at: https://www.tandfonline.com/doi/full/10.1080/21670811.2019.1623700

[73] Cadwalladr, C., & Graham-Harrison, E. (2018). The Cambridge Analytica files. *The Guardian*, 21, 6-7.

[74] Moore, M., & Tambini, D. (2018). *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*. Oxford University Press.

[75] Helberger et al., 2018

[76] Lewis, P. & McCormick, E. (2018). How an ex-YouTube insider investigated its secret algorithm. *The Guardian*, February 2, 2018.

[77] Lewis, 2018

[78] Tufekci, 2018

[79] Lewis, 2018

[80] Kaiser, J. & Rauchfleisch, A. (2019) The implications of venturing down the rabbit hole, *Internet Policy Review*, June 27 2019

[81] Serrato, R. (2018) How YouTube's algorithm amplified the right during Chemnitz, Algorithmic Accountability Reporting at AlgorithmWatch, Berlin, November 5 2018.

[82] Bahara et al., 2019

[83] Gary & Soltani, 2019

[84] European Commission (2018). *Code of Practice on Disinformation.* Available at: https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

[85] Council of Europe, 2018. Recommendation CM/Rec(2018)1[1] of the Committee of Ministers to member States on media pluralism and transparency of media ownership. Available at https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680790e13.

[86] Mosseri, A. (2018). Bringing People Close Together. *Facebook Newsroom.* Available at: https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/; Facebook (2018). How People Help Fight False News. *Facebook Newsroom.* Available at: https://newsroom.fb.com/news/2018/06/inside-feed-how-people-help-fight-false-news/; Zuckerberg, M. (2018). A Blueprint for Content Governance and Enforcement. *Facebook Notes*. Available at: https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

[87] Bickert, M. (2019). Combatting Vaccine Misinformation. *Facebook Newsroom*. Available at: https://newsroom.fb.com/news/2019/03/combatting-vaccine-misinformation/.

[88] Rosen, G. (2019). Remove, Reduce, Inform: New Steps to Manage Problematic Content. *Facebook Newsroom*. Available at: https://newsroom.fb.com/news/2019/04/remove-reduce-inform-new-steps/.

[89] YouTube, 2019

[90] Helberger, Leerssen & Van Drunen (2019). Germany proposes Europe's first diversity rules for social media platforms. *LSE Media Policy Project Blog.* Available at: https://blogs.lse.ac.uk/mediapolicyproject/2019/05/29/germany-proposes-europes-first-diversity-rules-for-social-media-platforms/.

[91]https://www.staatscommissieparlementairstelsel.nl/documenten/rapporten/samenvattingen/12/13/eindrapport

[92] https://www.vox.com/2019/6/26/18691528/section-230-josh-hawley-conservatism-twitter-facebook

[93] Cobbe & Singh, 2019; Gary & Soltani, 2019

[94] Ibid.

[95] Council of Europe, 2018, par. 2.5.

[96] Van Drunen, Helberger & Bastian (2019). Know your algorithm: what media organizations need to explain to their users about news personalization. *International Data Privacy Law* 9(3). https://academic.oup.com/idpl/advance-article/doi/10.1093/idpl/ipz011/5544759.

[97] MacCarthy (2020). Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry.

[98] Harambam, J. (2017) "The Truth Is Out There": Conspiracy culture in an age of epistemic instability. Rotterdam: Erasmus University.

[99] Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.; Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 1-22.; Swire et al., 2017; Lodge, M., & Taber, C. S. (2013). *The rationalizing voter*. Cambridge University Press.

[100] Graves, L. (2018). Understanding the promise and limits of automated fact-checking. *Factsheet*, *2*, 2018-02.

[101] Graves, L. (2016).

[102] Harambam, J. (2017b). De/politisering van de Waarheid. *Sociologie*, *13*(1), 73-92.

[103] Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*(6), 701-723.; Lewandowksy et al., 2017

[104] Strandberg, K., Himmelroos, S., & Grönlund, K. (2019). Do discussions in like-minded groups necessarily lead to more extreme opinions? Deliberative democracy and group polarization. *International Political Science Review*, *40*(1), 41-57.; Wojcieszak, M. (2011). Deliberation and attitude polarization. *Journal of Communication*, *61*(4), 596-617.

[105] Cobbe & Singh, 2019

[106] Id.

[107] Edwards, L., & Veale, M. (2017) Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review*, *16*, 18.

[108] The technical challenge, however, of anonymizing datasets should not be underestimated. *See* Elizabeth Gibney, Privacy hurdles thwart Facebook democracy research https://www.nature.com/articles/d41586-019-02966-x?sf220739510=1. *See also* Kobbi Nissim et al., Differential Privacy: A Primer for a Non-technical Audience, https://privacytools.seas.harvard.edu/files/privacytools/files/nissim_et_al_-_differential_privacy_primer_for_non-technical_audiences_1.pdf

[109] Caltrider, J. Journey From the Dark Side: How One Teen Boy Got Radicalized Online and Came Out the Other Side. *Mozilla Foundation* (2019, August). https://foundation.mozilla.org/en/blog/journey-dark-side/

[110] Harambam, J., Helberger, N., & van Hoboken, J. (2018). Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180088.

[111] Harambam, J., Bountouridis, D., Makhortykh, M., & Van Hoboken, J. (Sept. 2019). Designing for the better by taking users into account: a qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems* (pp. 69-77). ACM.

[112] Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The information society*, *34*(1), 1-14.

[113] Committee of Ministers Recommendation CM/Rec (2018)1 on media pluralism and transparency of media ownership. Council of Europe 2018.

[114] Helberger, et al., 2018

[115] *See* McCarthy, Mark (2020). Transparency Requirements for Digital Social Media Platforms Survey and Recommendations for Policy Makers and Industry.

[116] Gary & Soltani, 2019

TRANSATLANTIC WORKING GROUP

Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective†

Heidi Tworek, University of British Columbia[1]

Ronan Ó Fathaigh, Institute for Information Law, University of Amsterdam[2]

Lisanne Bruggeman, Institute for Information Law, University of Amsterdam[3]

Chris Tenove, University of British Columbia[4]

January 14, 2020

Contents

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

Introduction

In May 2019, Instagram announced that users could now appeal if they disagreed with the company's decision to remove a post.[5] Let that sink in – until a few months ago, users could not appeal takedowns on one of the largest social media sites. Even now, those decisions are made internally with the appeal reviewed by a second content moderator.

Instagram's decision followed its parent company, Facebook, which had already allowed appeals. All these appeals rely upon trusting the legitimacy of private company processes. For any dispute resolution to function, everyone involved should feel that the dispute resolution mechanism holds legitimacy. That principle of legitimacy underpins our judicial systems and processes. It enables citizens to accept the results of court decisions and arbitrations. At present, social media companies do not command the same legitimacy. Many argue that they never can, nor should they as private entities.

Many users, scholars, civil society organizations, and policy makers have lost faith in social media companies' content moderation systems. The Electronic Frontier Foundation has described these systems as "fundamentally broken."[6] Problems abound: lack of respect for human rights standards, due process, and transparency; vague rules; inconsistent policy enforcement.

This paper offers suggestions for how to improve one specific aspect of content moderation: resolving disputes over takedowns. Dispute resolution may seem like a small area, but it actually encapsulates the legitimacy problems behind content moderation decisions. The mechanisms behind the decisions seem arbitrary to outside observers. Decisions are subject to no public scrutiny or accountability. Appeals can only be directed to the companies themselves, if appeals mechanisms even exist.

This paper considers the problem of legitimacy in content moderation decisions and suggests new institutions to rebalance the private and public interests in resolving disputes. David Kaye, the U.N. Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, has noted that appeal mechanisms and remedies are "limited or untimely to the point of nonexistence."[7] He has recommended that "[g]iven their impact on the public sphere," social media companies "open themselves up to public accountability."[8] Our recommendations thus contribute to a broader conversation about rebalancing the relationship between companies, users, and governments. This is otherwise known as platform governance.[9] Following the particular concerns of the Transatlantic Working Group, we seek solutions that safeguard freedom of expression.

Users seem to value companies' internal appeal mechanisms. In the first three months of 2019 alone, users challenged Facebook's removal of over 1.1 million pieces of content under its own rules on "hate speech."[10] Facebook restored just 10 percent of this content. There is no further public mechanism to dispute the decision. These numbers only cover users who took the time to complain, and only about hate speech, with millions more pieces of content removed under propaganda, violence, and harassment rules. The number of appeals raises the question of whether we need new institutions to adjudicate some of these decisions. This would help the companies, too, because currently they find themselves constantly under fire for their content moderation.

This paper makes two interventions. First, we propose five basic principles to bolster legitimacy in content moderation decisions: fairness, accountability, independence, transparency, and effectiveness (FAITE). Second, we suggest two possible institutions to achieve these principles in dispute resolution. First, social media councils could discuss terms of service, adjudication processes, and broader ethical questions. A second, complementary institution is e-courts.

The paper first introduces dispute resolution and content moderation. Section 2 analyses social media councils; Section 3 explores e-courts and online dispute resolution initiatives. Sections 2 and 3 draw on case studies from Europe and North America. The fourth section concludes with policy options and recommendations. The appendix provides further details on the case studies.

This paper is authored by Heidi Tworek, University of British Columbia (project lead); Ronan Ó Fathaigh, Institute for Information Law, University of Amsterdam; Lisanne Bruggeman, Information Law and Policy Lab (ILP Lab), Institute for Information Law, University of Amsterdam; and Chris Tenove, University of British Columbia. Nico van Eijk (Institute for Information Law, University of Amsterdam) was involved in the drafting phase until November 1, 2019. We are grateful to the members of the Transatlantic Working Group for their comments on various drafts of this paper before, at, and after our meeting in Bellagio in November 2019.

1. Key questions

- What principles should undergird dispute resolution mechanisms?
- How can we create legitimate dispute resolution mechanisms for content moderation?
- How do we strike a balance between privatized and public adjudication of online content?
- How do we better align content moderation with international human rights law and fundamental principles of freedom of expression?
- How do we bake freedom of expression into any institutions involved in social media?
- How do new institutions deal with the scale of online content?
- Who will finance new institutions?
- How do we ensure access to justice for content moderation questions?
- How do we create procedural safeguards to avoid any new systems being gamed?
- Who should be involved in discussing content moderation decisions?
- How should we deal with jurisdiction for content moderation?
- Who should be allowed to flag or complain about content moderation decisions?

2. Content moderation and dispute resolution mechanisms

Almost every day, new research documents the problems with social media companies' content moderation systems. These problems are acute for freedom of expression. One recent example was

that Google's AI tool to detect toxic comments often classified comments in African-American English as toxic.[11] Other minority and marginalised groups around the world fear suppression of their speech. Racist and misogynist content, often including death threats, may not be removed. At the same time, automated tools may amplify over-blocking and pre-publication censorship of political expression. And all of this is based on vague rules, which create chilling effects on discussing matters of public concern, like sexual health.[12] These are real-world harms, preventing individuals from exercising the most protected form of freedom of expression under international standards: political expression and expression on matters of public interest.

The time has arrived for democratic governments to play a more active role. The online environment has already become the main (and unprecedentedly open) space for exercising the right to freedom of expression. This means that the online world too is subject to international and regional human rights standards for freedom of expression. These standards stipulate that governments should not interfere with freedom of expression. Simultaneously, governments have a positive duty to enable a favourable environment for freedom of expression.[13] This is particularly true for minority opinions and ideas, and those considered offensive and shocking by government authorities and the majority. Many democratic governments also bear responsibility to promote diversity and pluralism and to promote equitable access to the means of communication. This duty extends to protecting relations between individuals and private companies.

We argue that democratic governments' more active role extends to creating institutions specifically for dispute resolution. We see dispute resolution as fundamental for a legitimate regulatory framework for the online environment.[14] We cannot assume that we can know what types of content are being deleted. Instead of focusing on particular types of content, we suggest institutions that can empower users. Social media were often initially praised as giving a voice to the voiceless. Yet AI tools used in content moderation now may disproportionately affect marginalised communities, including people of colour and women.[15]

Our paper suggests that e-courts could enable users to resolve disputes over content deletion through public scrutiny, when appropriate. Social media councils, in turn, could make content moderation more transparent, encourage information-sharing from companies, and create baseline standards for content moderation that safeguard freedom of expression and human rights.

Within dispute resolution, our paper mainly addresses the following scenarios:

1. A user's content has been removed, or a user's account has been suspended, based on a supposed violation of the social media company's terms of service, adopted community guidelines, or within the context of legally cognizable rights. The removal or suspension may have resulted from the company's own proactive review of content or a complaint (flagging) by another user or third party (e.g., civil society organisations).

2. A complaint that another user's content *should be* removed, but has not been removed. The complaint may have arisen from another user (e.g., a journalist) or a third party, such as a civil society organisation, and concerns content that is not removed despite being problematic.

3. Essential standards

Before exploring the possible models, we suggest essential criteria or standards for dispute resolution mechanisms. We derive these standards from the wealth of literature on independent self-regulatory bodies, particularly independent media regulatory authorities. We draw inspiration from legal instruments which set out criteria to guarantee the independence of national regulatory authorities from government and private interests.[16] We see five fundamental principles (FAITE):

1. *Fairness:* Dispute resolution bodies must have fair procedures. All parties (users, social media companies, civil society organisations and other institutions/stakeholders) must be able to express their point of view, be provided with arguments and evidence put forward by other parties, and be able to comment on them. Dispute resolution bodies must have the necessary expertise, including members possessing the necessary knowledge and skills in the fields of freedom of expression and international human rights law.

2. *Accountability*: Dispute resolution bodies must be accountable to the public and civil society; all decisions taken and regulations adopted should be duly reasoned.

3. *Independence*: Dispute resolution bodies must be as independent as possible or their membership must be balanced to try to ensure independence. Members should hold terms of office long enough to ensure the independence of their actions. Members should disclose any circumstances that may, or may appear to, affect their independence or create a conflict of interest.[17]

4. *Transparency*: Any bodies should be established via a fully consultative and inclusive process. Their work should occur in an open, transparent, and participatory manner with transparent procedural and financial rules and decision-making procedures. Members must be appointed through a transparent procedure. Decisions must be published in easily accessible formats.

5. *Effectiveness*: Dispute resolution bodies must have adequate financial and human resources to carry out their function effectively; any complaint procedure should be free of charge and easily accessible. Dispute resolution bodies must have adequate powers to provide appropriate remedies.

4. Range of regulatory measures

Dispute resolution has long existed. Many industries have struggled with similar questions for decades, including the press, broadcasting, and advertising. We can thus draw on a wide range of self-regulatory, co-regulatory and statutory regulation models for dispute resolution.

First, there is **self-regulation**, which can take many forms.[18] It can include self-regulation through *internal governance*, where companies themselves establish policies based on their own cultures, practices and control. Content moderation policies are usually developed by a social media company's legal officials, public policy and product managers, and senior executives.[19] However, companies may also engage *external* parties to a certain degree, but still ultimately subject to the company's own governance structure. Examples include Twitter's Trust and Safety Council, which provides expert input on products, policies, and programmes;[20] and Google's short-lived Advanced Technology External Advisory Council, to monitor responsible development and use of AI.[21] Facebook has also announced a planned Oversight Board.[22]

This type of self-regulation can be complemented by self-regulation through *user regulation*, where users flag breaches of content rules, such as YouTube's Trusted Flagger programme, assisting with enforcement of YouTube's community guidelines.[23] Another example is Wikipedia's Arbitration Committee, known as Wikipedia's supreme court, which considers serious conduct disputes and appeals over blocked or banned users.[24]

Further, there is self-regulation through *industry standards* and codes of conduct. The Global Network Initiative, for example, seeks to protect freedom of expression and privacy in the ICT industry by setting a global standard for responsible company decision-making.[25] The Internet Watch Foundation, established by the internet industry, reports potentially criminal child sexual abuse content online.[26] Crucially, this raises the question of the independence of regulatory bodies established by industry.[27]

Second, there is **co-regulation**, which can also take many forms, and encompasses self-regulation plus an element of regulatory compulsion to enable effective enforcement. Government, legislation, or a statutory regulator can recognise self-regulatory bodies. Alternatively, government or regulators can approve codes of conduct, for example with the EU's Business-to-Platform Regulation, which regulates disputes between certain online platforms and business users. While the legislation does not mandate that online platforms must resolve disputes through independent mediation, it sets out a list of statutory requirements (e.g., impartiality and independence) for any such alternative dispute resolution.

Finally, there is **statutory regulation**, where legislation sets the scope and coverage of regulation, which may be enforced by a statutory regulator. This can prove problematic and undermine freedom of expression. One example is Russia's 2019 amendments to the Information, Information Technologies and Protection of Information Law, which provides that the communications regulator may order the removal of illegal online content without a court order, including "unreliable socially significant information."[28]

In the sections that follow, we discuss two potential and complementary options that adhere to international freedom of expression standards, and provide fair, independent, accountable, transparent, and effective dispute resolution mechanisms for content moderation.

Social Media Councils

In this section, we first discuss current proposals for social media councils, and some industry initiatives. We then delve into fundamental questions around social media councils.

1. Current situation

The idea of social media councils has gathered significant steam in the last two years. The U.N. Special Rapporteur described independent social media councils as "[a]mong the best ideas," noting that "[e]ffective and rights-respecting press councils worldwide provide a model for imposing minimum levels of consistency, transparency and accountability to commercial content moderation."[29] ARTICLE 19 put forward the idea of independent social media councils in early 2018.[30] Tenove, Tworek, and McKelvey proposed content moderation councils for Canada.[31] In late 2018, Facebook announced the idea of its own oversight board.[32] In early 2019, Stanford University's Global Digital

Policy Incubator (GDPi), ARTICLE 19, and the U.N. Special Rapporteur held a multistakeholder conference on social media councils.[33]

There are two main models for independent social media councils. One (the GDPi model) would begin from the global level, designed to provide guidelines and evaluate emblematic cases, not adjudicate individual complaints. The second (ARTICLE 19), a national or regional council, would allow users to bring a claim, but the council has discretion over its docket.

The GDPi model would be a global multistakeholder cross-platform social media council, designed to complement company-specific and national initiatives. The council would develop guidelines based on international human rights principles for how to approach content moderation online. It would *not* adjudicate individual cases or serve as an appeals body. Rather it would evaluate emblematic cases and create precedents for future cases. Further, it would advise platforms in developing and implementing their terms of service, mediate the interaction between governments and platforms, and provide advice, recommendations, and expertise to governments.

ARTICLE 19's model would be a network of national or regional social media councils, providing general guidance to social media platforms and, crucially, deciding individual users' complaints. National and regional councils would apply international standards on human rights, either directly or on the basis of a universal Code of Principles adopted globally. The social media councils would have an adjudicatory function. Individual users would have the right to bring a complaint, though only after adjudication at the platform level to avoid the risk of overload. A social media council would have full control of its docket, and could filter the cases it wants to review to "build a relationship of trust with the public."[34] In summer 2019, ARTICLE 19 published a consultation paper on social media councils as another phase in developing the idea.[35] These national councils could establish working relationships with any global body.

Social media companies, whether in cooperation or individually, are also establishing procedures for content moderation. An example of industry cooperation is the Global Internet Forum to Counter Terrorism (GIFCT), established in 2017 by Facebook, Twitter, YouTube and Microsoft. It enables coordination between social media companies seeking to remove "terrorist" content, and the GIFCT houses a hash database of "terrorist" images. However, as Tworek notes, there is no public oversight, and it remains a "mystery to those outside the companies or specific civil society organizations and governments who cooperate with the forum."[36] At a meeting between governments and technology companies at the U.N. General Assembly in September 2019, as part of the Christchurch Call to Action to eliminate terrorist content online, structural changes to GIFCT were proposed, including a government oversight board.[37] The GIFCT is now transforming into a nonprofit which may entail more transparency and ability for public oversight.

There are also company-specific plans. In September 2019, Facebook published further details of its planned oversight board.[38] Facebook proposes to establish an oversight board comprising 40 members who serve three-year terms. Facebook will select the co-chairs. Then, Facebook and the co-chairs select the rest of the members.[39] The board will review and decide on content according to Facebook's "content policies and values." If people disagree with the outcome of Facebook's decision and have exhausted their appeals, they can submit a request for review to the board.[40] The board chooses which requests to review and adjudicate. The board's resolution "will be binding and Facebook will

implement it promptly, unless implementation of a resolution could violate the law."[41] Facebook plans to start the system in 2020.[42]

Facebook's plans show that the company has recognised the problem of legitimacy in internal appeals. But questions remain over implementation and the value of a company-specific process. Social media councils offer a potential avenue to address these concerns. Below we consider some fundamental questions on how they might work.

2. Fundamental questions on social media councils

a. *Geography*: Should social media councils be national or international?

The geographical coverage of a social media council is central to all its potential functions, composition, and regulatory frameworks. The GDPi and ARTICLE 19 models offer two different approaches to this question, one international and one national. We suggest that a national social media council may be better placed to address three of the major concerns with the current system of content moderation and complaint-handling by social media companies: insufficient understanding of linguistic and cultural nuance; absence of rigorous human evaluation of context, including wholesale problems with addressing cultural context; and an inability to understand widespread variation of language cues, meaning, linguistic and cultural particularities.

This view is consistent with European press councils, which apply national codes of ethics and often deal with international media conglomerates. The Alliance of Independent Press Councils of Europe (a network of European press and broadcast councils) sees universal codes of ethics as impossible. Although press and media councils are national, almost all include international media firms. National councils and international companies could similarly cooperate on social media.

Further, national councils could complement each other in the transatlantic context, given the different freedom of expression standards. Countries could coordinate to create similar governance structures and composition. A national framework also recognises that different countries are at different stages of thinking on this issue. Some wish to move far faster than others, making international coordination between national councils more realistic at this point.

b. *Scope*: Should social media councils decide guidelines or examine individual cases?

Social media companies' terms of service emerged on an ad hoc basis. Although these terms of service are now much more systematised, they did not place freedom of expression standards at their core. Ironically for companies generally built on the philosophy that more speech was better, there are now grave concerns that content moderation decisions actually harm freedom of expression. Because the basic problem emerged from arbitrary standards, a social media council could start by creating guidelines that embed freedom of expression standards in social media companies' content rules.

But guideline-making would only go so far. The true test of guidelines is implementation and accountability. That is why the press industry established press councils *in addition* to codes of ethics. A social media council limited to guideline-making would leave implementation to social media companies themselves. It would also leave users without a crucial right to subject a social media company's decision to independent review. Indeed, freedom of expression is so important that it

should include the procedural safeguard of review by an independent and impartial body. A social media council limited to guidelines-making would not offer such a procedural safeguard. One solution would be to send individual cases to e-courts.

Further, a social media council limited to guideline-making would not solve the crucial problems of (a) widespread inconsistency and unpredictability in content moderation decisions; (b) inability to challenge or follow-up on content-related complaints; (c) lack of transparency and due process; (d) widespread use of automated decision-making; and (e) insufficient remedies.

Alongside guidelines, social media councils could help to coordinate third-party research into companies, compile decisions from different companies, and compare their content moderation policies. They could create industry-wide boards of ethics for new questions like AI and develop best practices for crucial questions like labour standards for content moderators.[43]

Finally, a decision-making mechanism could cure the major deficiencies in the remedies available. Current systems offer no publication of a reasoned decision, no right of reply, no publication of an apology or acknowledgment. Where content removal causes specific reputational, physical, moral and/or financial harm, there are few mechanisms for settlements. A national social media council with a decision-making function could provide some of these remedies. Social media councils could develop and apply national standards, akin to how press councils make decisions based on national press codes. Here, civil society engagement could ensure that social media companies' content moderation rules and systems reflect freedom of expression principles.

c. *Regulatory framework*: What type of body might a social media council resemble?

There are myriad regulatory frameworks spanning from self-regulation to statutory regulation. With social media councils, the major question is whether they should be voluntary. The voluntary nature of social media councils may help build trust and accountability between the public and social media companies. However, if social media companies are unwilling to send representatives, governments may start to contemplate other regulatory options. Canada, for example, offers three possible frameworks (see the appendix for more detail on each).

i) Self-regulation like the National NewsMedia Council or an Ads Standards Council.

It is tempting to frame social media councils as updated media councils. We cannot forget, however, the fundamental differences between user-generated social media content and journalism produced by a comparatively small number of newspaper and broadcast professionals. In both cases, relatively few companies are involved. But the content producers are very different – journalists versus almost every member of the public. The scale differs vastly too. Any social media council would need to account for these fundamental differences.

ii) Co-regulation like a Broadcasting Standards Council.

Private radio and television broadcasters created the Canadian Broadcasting Standards Council (CBSC) to devise industry codes and to address audience complaints about their programming, including when it is streamed online. Membership in the CBSC is voluntary, but non-members have complaints addressed directly by the government regulator of the entire broadcasting sector (the

Canadian Radio-television and Telecommunications Commission, or CRTC). The CRTC functions as an appellate body for those unsatisfied by CBSC judgments.

A social media council based on this framework would help develop and implement standards in a largely self-regulatory manner, while involving stakeholder input and government oversight. Consultations to create the council must include major national and international internet companies, and government must offer appropriate incentives – or threats of more intrusive regulation – to secure participation. To prevent the council's composition becoming lopsided, the participation of other key stakeholders would be encouraged and supported, ranging from Indigenous communities to human rights organizations to political parties.

A social media council constituted under the auspices of the CRTC would apply a new group-based approach, pursuing "broadly based agreements tailored to and established with a few dozen specific companies or affiliated groups of companies, individually or collectively offering a variety of services (service groups)" in order to provide "public scrutiny and should set out specific binding commitments applicable to the service group."[44] While we caution against using this experimental framework for online broadcasting or digital common carriers, the content moderation industry could fit as a group in this new "nimble regulatory" approach. The social media council would then be responsible for enforcing the group's service agreements.

iii) A human rights framework and human rights bodies.

At the federal and provincial levels, human rights legislation sets out means to address rights. Human rights bodies could provide guidance and act as a complaints body to address violations of certain rights through the use of social media platforms. Human rights bodies have issued decisions regarding hate speech and discrimination on websites and message boards. More recently, the Canadian and Ontario human rights commissions jointly demanded that Facebook introduce safeguards to protect against discrimination in targeted employment advertising.[45] The Canadian Human Rights Commission (CHRC) is most likely to have jurisdiction over issues related to social media content moderation.

Nevertheless, the CHRC faces several challenges in addressing content moderation issues. First, current human rights legislation applies to a very narrow set of issues, particularly since the federal Conservative government removed the provision addressing hate speech in the Canadian Human Rights Act in 2013. Second, there are complex jurisdictional questions regarding the application of federal or provincial human rights frameworks. Different bodies have jurisdiction over different issue areas (e.g., employment, hate speech, etc.), and subject entities (e.g., broadcasters, web publishers, etc.). The CHRC's decisions are enforceable at the national level, but the CHRC's framework may be coordinated globally via the Global Alliance for National Human Rights Institutions and through reference to international laws and guidelines. Greater clarity would be necessary, either through new legislation or precedent-setting decisions.

d. *Composition*: Who should sit on a social media council?

Composition is a complicated and crucial question. As with European press councils, each country may take a different approach. The Netherlands Press Council, for example, has a chairperson, four

vice-chairs, 10 member-journalists, and 10 non-journalist members. In Germany, however, only publishers and journalists comprise the Press Council board with no independent representatives.

i) How should civil society be involved?

Obviously, not every social media user can have the right to sit on a social media council. But different groups of users need representation for a council to gain any public legitimacy. Civil society involvement is crucial, unlike current press or media councils. Civil society involvement in creating social media councils, and broad representation, is vital for public trust. Civil society could possibly lodge complaints with councils, support complaints, and lodge third-party submissions. Finally, civil society would operate as an important check on the operation of social media councils, and assess whether they are contributing to an enabling environment for freedom of expression online. Each country would have to consider carefully and consult about which civil society organisations or representatives might contribute to the council. The council might particularly try to include those who have suffered most from hate speech or abusive speech.

ii) What size of social media company should participate?

In policy circles, "social media companies" is often used as a synonym for Facebook and Twitter. But this overlooks influential alternative networks such as Reddit, or Mumsnet in the UK. Given that participating in councils can involve significant company time, we should be attentive to the potential burdens of membership. One approach is only to include companies over a certain size. Another is to apply a sliding scale of participation based on size, market capitalisation, or another measure.

Another approach is to think about the council starting from the perspective of smaller companies. Any initiatives in this space should be wary of moves that unintentionally "lock in" the big players and stifle smaller companies. European press councils, for example, often involve not only dominant and large pan-European media companies, such as Axel Springer or Sanoma, but also small online-only news publications, with limited readership and staff.

e. *Diversity of approaches*: Is it valuable to maintain multiple approaches to content moderation?

Smaller and medium-size companies often bring very different approaches to content moderation. Even the major companies differ significantly, as the recent policy decisions on political advertising show. This raises a broader concern about whether social media councils should support a diversity of approaches to content moderation or create more uniform standards.

At base, this is a question about competition: do we want social media councils to encourage uniformity or do we want an ecosystem where different social media companies make different decisions? There are advantages both to uniformity and multiplicity (especially in a truly competitive environment with adequate substitutes). One way to thread the needle is for social media councils to create baseline standards for freedom of expression. Above that baseline, companies could implement their own content moderation policies. This could address the widespread inconsistency and unpredictability in content moderation decisions.

Rather than imposing uniform content moderation, a social media council can create uniform consistency and predictability of content moderation decisions.

f. *Jurisdiction.*

As with any online dispute, jurisdiction is a major consideration. Here, we suggest it could depend upon the type of dispute. We can look at this through the two different types of disputes we sought to address in this paper.

i) The first category is content removed or an account suspended based on a supposed violation of the social media company's own content rules or terms of service. The removal or suspension may have resulted from the company's own proactive review of content or having been flagged by another user (e.g., journalist) or third party (e.g., civil society organisations). Here, jurisdiction could be based upon the location of the user whose content or account was flagged.

ii) The second category is a complaint that another user's content should be removed, based on the company's own content rules or terms of service, but it is not. Here, jurisdiction is trickier. The complaint may have arisen from another user, or third party, such as a civil society organisation or government agency, based in another country from the content's author. If social media councils allow these types of public-interest complaints where complainants are not directly affected, then jurisdiction could depend upon the location of the content's author.

Most realistically, social media councils could facilitate discussion with civil society, create baseline standards and ethics, and improve company transparency. We suggest that individual appeals about content removal might best be directed to a second new institution: e-courts.

E-Courts

A second possible dispute resolution model is e-courts or internet courts. E-courts exist throughout the world to resolve disputes between governments and individuals, companies and individuals, disputes between individuals, and disputes between companies. In the Netherlands, for example, various types of e-courts started in 2009. However, there is no agreed definition of e-courts, and they take many forms. As one legal scholar notes, online dispute resolution bodies "generally facilitate settlement or substantive determination on the merits," while e-courts are "more limited to ending the dispute or providing a remedy or result based on limited parameters."[46] For content moderation, e-courts could be crucial to upholding freedom of expression standards.

Dispute resolution is already fundamental for many online companies. One of the best-known examples is eBay's online dispute resolution mechanisms, which resolve over 60 million disputes between eBay traders and users annually.[47] One programme is for feedback disputes, where an eBay seller may challenge a review posting about a seller's businesses. As part of eBay's online dispute resolution procedure, an impartial third-party reviewer from a professional dispute resolution service can examine the challenged posting and determine whether to affirm, withdraw or take no action on the review.[48] Here, we consider how to create public e-courts.

1. Fundamental questions on e-courts

a. *Regulatory model:* What is a feasible regulatory model for dispute resolution of content?

There are three main appropriate models for e-courts. Unsurprisingly, the feasibility of e-courts fundamentally depends upon the type of e-court. Both Europe and North America have already created myriad e-courts which can help us to understand what has worked and what has not.

i) Online judicial adjudication.

This model uses technology to bring speed, efficiency, and lower costs. These are court-operated and judges make decisions. The model includes public or private online platforms for electronic case management, filings and communications. Examples include the Michigan Online Case Review in the United States and the European Small Claims Procedure, both discussed in further detail in the appendix. These e-courts digitise court processes such as small civil claims or minor offences, using technology to bring speed, efficiency, and lower costs.

Online judicial adjudication would provide a fast, simple, and cheap mechanism for users and others to challenge content moderation decisions made by social media companies. It would be fully online with no physical presence of the parties. Specially trained judges would make decisions. The simplified procedure would resemble small claim procedures; there would be no right of appeal to the general courts. Social media companies would assign specialised lawyers to process claims before the e-courts. The e-courts would regularly publish case-law compilations. Finally, the social media company would have to acquiesce to e-court jurisdiction to operate in a state.

Notably, this model would probably require legislation. The European Small Claims Procedure required legislation to be enacted. An e-court modelled on such a procedure at the European or national level would seem to require legislation – particularly if there were no right of appeal to general courts, and social media companies were subject to mandatory jurisdiction and the rules in order to operate.

ii) Online dispute resolution with built-in independent adjudication.

This model encourages early resolution through online dispute resolution, while preserving independent adjudication as a last resort. The model goes beyond digitising court systems, and involves online negotiation, independent mediation, arbitration, and adjudication. One of the best examples is the British Columbia Civil Resolution Tribunal, with its tiered approach: (a) online solution platform; (b) online facilitated negotiation; (c) online facilitated settlement; and (d) online hearing and adjudication.

Similar to the first model, legislation would be required to establish such an e-court, as it would include rules on access to courts, and judicial review of decisions. This model is also specifically designed to encourage early dispute resolution, and adjudication only as a matter of last resort, and therefore may save resources. Further, the independent adjudication would be carried out by officials established by legislation, rather than judges. This raises the question of whether this model is akin to statutory regulation, as the Civil Resolution Tribunal acts similarly to a statutory regulator, like a broadcasting regulator in Europe.

iii) Online independent dispute resolution, which may be operated by both private and public bodies.

This has a similar procedure to the previous model, but without independent adjudication. The best example operated by a public body is the European Online Dispute Resolution platform, and would involve: (a) online solution platform; (b) online facilitated negotiation; and (c) online facilitated settlement. This model can be put on a legislative footing, but it is not necessary.

Unlike the other models, this process may be established by private companies themselves. This would be the most feasible model not requiring legislation, and could also be built to encourage early dispute resolution, rather than proceedings through courts. It would also offer users an independent out-of-court dispute resolution mechanism. Of course, such mechanisms would need to be consistent with the EU's Alternative Dispute Resolution Directive, and some social media companies do already offer similar procedures for business users.

b. *Scalability*: How can e-courts cope with the potential scale of cases?

The most obvious question about e-courts is whether the low barrier to entry and admirable principle of access could create an unmanageable caseload. One method to cope within the EU is to scale national e-courts up to a regional e-court, and take the European Small Claims Procedure and European Online Dispute Resolution as possible models for a network of national e-courts.

Further, extant e-courts suggest a manageable scale. In 2017, the European Online Dispute Resolution platform handled over 24,000 complaints, the Dutch Stichting e-Court handled more than 20,000 cases, and in its latest figures, the Civil Resolution Tribunal in Canada handled nearly 12,000 disputes. The Civil Resolution Tribunal has been gradually extending its competence beyond the initial small claims it handled, and could provide a model for how to scale an e-court with more competences. Finally, for a benchmark from a statutory regulator, the UK broadcasting regulator (Ofcom) dealt with over 55,000 complaints last year.[49]

c. *Accessibility*: How can e-courts ensure accessibility while preventing gaming of the system?

One major advantage of e-courts is that they can enable greater access to justice. With reduced costs and the ability to file from home, an e-court could lower the barrier to entry. It could also help to prevent some current speech-law cases that create a Streisand effect, meaning that a case unintentionally draws even more attention to the information or speech that the case tried to prevent.

A separate concern in the online environment is that malevolent actors may try to game the system and launch multiple complaints against a user's content to try to get that person removed or their content blocked. The e-court system could perhaps use automated detection of repeated submissions to prevent anyone coordinating to flood the system. More broadly, any institution would need to have procedural safeguards to prevent the court system being used as a tool of trolling or harassment, the very things it was designed to prevent.

d. *Compatibility*: How are e-courts compatible with European, American, or Canadian law?

Both North American and European legal systems have already given legislative footing to e-court models. At the European level, the European Online Dispute Resolution and European Small Claims Procedure provide a framework. The examples from the Netherlands in the appendix demonstrate the operation of national e-courts.

One crucial point about e-courts is that government legislation may be needed to create the institution. But judges would decide individual cases, not government agencies. This is critical to uphold international freedom of expression standards.

e. *Finances*: Who will pay for e-courts?

Like any judicial institution, public authorities would seem the obvious choice to cover the cost of e-courts. Their online nature should make them comparatively inexpensive. Another option is that platforms bear the cost of the service. There are currently discussions over tax policy related to online services. Any new or additional tax revenue from platforms could be used to pay for e-courts.

f. *Jurisdiction*.

This raises similar issues discussed in the previous section on social media councils, including whether individuals may make complaints where they are not directly affected. The e-court model would have to adopt rules on whether only individuals directly affected by content may initiate a claim, or whether civil society organisations, for example, may initiate group-based or public-interest claims. Although complex, these jurisdictional questions apply to similar issues like the right to be forgotten in the EU.

Conclusions and recommendations

This paper seeks to provide an appropriate framework for building dispute resolution mechanisms which are fair, accountable, independent, transparent, and effective. Having considered the range of social media council models in Section 2, and the range of e-court models in Section 3, we offer the following conclusions and recommendations:

1. *Fairness, accountability, independence, transparency and effective standards (FAITE) are essential*: Dispute resolution mechanisms for online content moderation must be consistent with the essential standards of fairness, accountability, independence, transparency and effectiveness. Both social media councils and e-courts can satisfy these standards, provided the standards are applied during the creation and operation of these bodies.

2. *The model adopted must connect to identified problems*: It is crucial to adopt a model that best remedies the problems associated with current content moderation, particularly legitimacy, and best provides an enabling environment for freedom of expression online. This paper addresses a subset of the most relevant problems in content moderation, such as when companies' terms of service/community guidelines or behaviour based on or following from other (legal) instruments may violate a person's fundamental rights.

3. *Dispute resolution mechanisms are a key part of establishing legitimacy and creating public accountability:* Independent complaint/dispute mechanisms can remedy myriad other problems of the current system operated by social media companies: widespread inconsistency and unpredictability in content moderation decisions; inability to challenge content actions or follow up on content-related complaints; lack of transparency, due process and notice; widespread problematic use of automated decision-making; and wholly insufficient remedies. Complaint/dispute mechanisms would provide the necessary public accountability of the freedom of expression standards implemented into social media

companies' content rules, and provide a crucial procedural safeguard of review by an independent and impartial body for users and the public.

4. *Social media councils and e-courts are two institutional options:* We have outlined two detailed models for institutions that could create greater public accountability and transparency. In an ideal world, these institutions can benefit everyone involved: users, civil society groups, companies, and governments. These two institutions will not solve all the problems of platform governance. But they will start toward finding a legitimate balance between the public and private interests in promoting freedom of expression.

5. *Social media councils and e-courts are not mutually exclusive:* Their compatibility depends upon the models for social media councils and e-courts. If a social media council is limited to guideline-making, an e-court would provide decisions in actual cases. If social media councils have a decision-making function, there could be coordination issues. A social media council could be combined with an e-court or social media councils could be involved in decisions preceding e-court decision-making.

6. *Freedom of expression should be a core value for any new institutions*: The current content moderation system has many potential detrimental effects on freedom of expression, which flow from the lack of freedom of expression standards built into social media companies' content rules. Freedom of expression principles do not form the basis of content moderation decisions. The opposite should hold for new democratic institutions like social media councils or e-courts. Freedom of expression concerns should be baked into the design, composition, and creation of any new institutions.

7. *Social media companies can help themselves by acting now:* Even without creating a social media council or e-court, social media companies themselves can already begin to alleviate these problems by incorporating freedom of expression standards into their content rules, and freedom of expression standards into their content moderation decisions and complaint-handling systems. Social media companies should invite relevant stakeholders to assist in this process, and formulate the appropriate freedom of expression standards upholding the FAITE standards of freedom of expression online.

Dispute resolution is complicated in every industry. But fair procedures are crucial to generate trust and legitimacy. This is even more true for such a sensitive area as speech. These two institutions of social media councils and e-courts are not panaceas for all the problems of platform governance. But they could go a long way in addressing some of the most pressing concerns.

Appendix

Here we provide deep dives on our particular case studies of councils and e-court systems. This material provides further details for those interested in the mechanics of particular systems.

Social Media Council Case Studies

1. Case study on social media councils in North America

For a North American case study, we focus on Canada. Canada does not currently regulate social media as a sector, though a range of existing regulatory frameworks apply. These include federal and provincial privacy regulation and regulation on advertising. They also include criminal, civil, administrative and human rights laws that address certain forms of prohibited speech, including defamation, hate propaganda, counselling terrorism and the non-consensual distribution of intimate images.

Multiple parliamentary committees and the Canadian government have recommended action to impose duties on social media companies "to remove manifestly illegal content in a timely fashion, including hate speech, harassment and disinformation, or risk monetary sanctions commensurate with the dominance and significance of the social platform, and allowing for judicial oversight of takedown decisions and a right of appeal."[50] The Liberal Party has formed a minority government following the federal election in October 2019. The party's platform claims that it will "target online hate speech, exploitation and harassment," and "when social media platforms are used to spread these harmful views, the platforms themselves must also be held accountable."[51]

Canada, as in many areas, pursues a regulatory framework that is partway between American and more interventionist European approaches. On the one hand, Canada has strong constitutional protections of freedom of expression, relatively minimal regulation of publishers and light regulation of broadcasters with respect to content. On the other hand, Canada has criminal laws against hate propaganda, and its human rights bodies have taken action on discriminatory or hateful communication disseminated by websites, publishers, broadcasters, and – in a few cases – social media companies.[52]

One report proposed a social media council as a policy measure.[53] Within Canada, there are three potential routes for a social media council: a self-regulatory body (similar to Canada's National NewsMedia Council or Ad Standards), a co-regulatory body facilitated by the Canadian Radio-television and Telecommunications Commission (similar to the Canadian Broadcast Standards Council), and a human rights approach (primarily pursued through the Canadian Human Rights Commission).

Self-regulatory social media council: Canada already has self-regulatory bodies for two cognate sectors to social media: the National NewsMedia Council (for news media other than broadcasters) and Ad Standards (the advertising sector). The National NewsMedia Council (NNC) is Canada's main press council. It is a voluntary, self-regulatory body for the news media industry, with over 500 member organizations. The council promotes media ethics and operates a process to hear complaints.

The Council does not have its own standards, but relies on "the news outlet's code of practice or generally accepted journalistic standards."[54] If a complaint process finds that the news organization failed to uphold these standards, and that adequate corrective measures were not taken, the organization can "publish a fair report including, at minimum, a summary of the decision, and a link to the decision on the NNC website." The NNC's enforcement capacities are very weak, leading to questions about its relevance.

Ad Standards is a self-regulatory body for Canada's advertising sector. It administers the *Code of Advertising Standards*, and operates a consumer complaints process about advertisements that run in Canadian media. The Code was created in 1963 after broad consultation with industry and stakeholders, and is regularly updated. Among its provisions, it prohibits advertisements that condone discrimination based upon race, national or ethnic origin, religion, gender identity, sex or sexual orientation, age or disability; that condone violence; that demean individuals or groups; or that "undermine human dignity." The Consumer Complaints Process encourages advertisers to take voluntary actions and address the complaints, and may forward the complaints to a Standards Council to make a decision. Ad Standards enforces a council's rulings by calling on media organizations to no longer host the advertising, and informing the Competition Bureau or other regulatory bodies that the advertiser is not compliant with the Code.

A co-regulatory social media council, under the auspices of the CRTC: A social media council could also follow the approach taken by private broadcasters in Canada. Private radio and television broadcasters created the Canadian Broadcasting Standards Council (CBSC) to address audience complaints about their programming, including when streamed online. Describing its mandate, the CBSC states:

> Broadcasters have the ability to influence opinion, modify attitudes and shape minds. That's why the industry created a voluntary system of codes that set high standards for all of their programming. Through these codes, private broadcasters promise to respect the interests and sensitivities of the people they serve, while meeting their responsibility to preserve the industry's creative, editorial and journalistic freedom.[55]

Membership in the CBSC is voluntary, but non-members will have complaints addressed directly by the government regulator of the entire broadcasting sector (the Canadian Radio-television and Telecommunications Commission, or CRTC), and the CRTC functions as an appellate body for those unsatisfied by CBSC judgements.

The CBSC applies industry codes on issues relating to ethics, violence on television, equitable portrayal, journalistic ethics, and cross-media ownership. When it receives complaints, the CBSC may constitute an Adjudicating Panel to investigate, arrive at, and publish a decision. If the broadcaster is found to have violated a code, it must announce that result several times on air. The CBSC cannot levy fines or revoke broadcast licenses.

A social media council based on this framework would help develop and implement standards in a largely self-regulatory manner, while involving stakeholder input and government oversight. Once established, the council could help social media platforms coordinate their actions to mitigate harmful speech, disinformation and other threats to democratic discourse online. Consultations to create the

council must include major national and international internet companies, and government must offer appropriate incentives – and threats of more intrusive regulation – to secure their participation. To ensure the composition of the council is not lopsided, the participation of other key stakeholders would be encouraged and supported, ranging from Indigenous communities to human rights organizations to political parties.

A social media council constituted under the auspices of the CRTC would apply a new group-based approach, pursuing "broadly based agreements tailored to and established with a few dozen specific companies or affiliated groups of companies, individually or collectively offering a variety of services (service groups)" in order to provide "public scrutiny and should set out specific binding commitments applicable to the service group."[56] While we caution against using this experimental framework for online broadcasting or digital common carriers, the content moderation industry could fit as a group in this new "nimble regulatory" approach. The social media council would then be responsible for enforcing the group's service agreements.

A human rights framework: Canada also addresses disputes over mediated communication through a human rights framework and human rights bodies. This offers an alternative or a productive complement to a social media council. At the federal and provincial levels, human rights legislation sets out means to address rights violations. Human rights bodies have issued decisions regarding hate speech and discrimination on websites and message boards. More recently, the Canadian and Ontario human rights commissions jointly demanded that Facebook introduce safeguards to protect against discrimination in targeted employment advertising.[57] The Canadian Human Rights Commission (CHRC) is most likely to have jurisdiction over issues related to social media content moderation.

The CHRC and some provincial human rights commissions bodies can constitute tribunals to hear and investigate specific complaints, compel evidence and enforce remedies. This human rights approach offers several advantages:

- The CHRC and other human rights bodies can constitute tribunals to hear and investigate specific complaints (from individuals or groups), compel evidence, and enforce remedies. Complaints may address those responsible for creating or disseminating problematic communications. These investigative processes follow stronger rule of law provisions than self-regulatory bodies, while generally being more accessible to complainants than civil or criminal processes.

- The CHRC and other human rights bodies can issue remedies that are legally enforceable. These can include retractions or removals of content and fines. Failure to adhere to decisions can lead to criminal prosecution.

- In addition to addressing individual complaints, the CHRC can study human rights issues, including systemic risks to rights. These studies enable the CHRC to comment on legislative proposals and play a public education role.

- Human rights bodies are bound by a commitment to freedom of expression as a constitutionally protected right of Canadians, and restrictions on speech through the rulings of Canadian human rights bodies can be reviewed by court systems.

- The Canadian Human Rights Commission applies a Canadian legal framework. However, it is a member of the Global Alliance for National Human Rights Institutions, which helps state-based human rights institutions to achieve shared goals and engage with the United Nations' Human Rights Council and Treaty Bodies. National human rights bodies may therefore be able to take a shared approach on issues related to social media content moderation, drawing on international human rights law and bodies.

Nevertheless, the CHRC faces several challenges in addressing content moderation issues that may affect the enjoyment of human rights. There are complex jurisdictional questions regarding the application of federal or provincial human rights frameworks. Different bodies have jurisdiction over different issue areas (e.g., employment, hate speech, etc.), and subject entities (e.g., broadcasters, web publishers, etc.). Greater clarity is needed, either through new legislation or precedent-setting decisions. Another challenge is that the provision addressing hate speech in the Canadian Human Rights Act was removed in 2013 by the federal Conservative government.[58] Policy makers and experts in Canada have argued for a new provision to address hate speech, but no concrete provision has been put forward to Parliament.[59]

Human rights bodies could provide guidance and act as a complaints body to address violations of certain rights through the use of social media platforms. Decisions by the CHRC are enforceable at the national level, but the framework applied by the CHRC may be coordinated globally via the Global Alliance for National Human Rights Institutions and through reference to international laws and guidelines. However, there remain significant jurisdictional challenges, and current human rights legislation applies to a very narrow set of issue areas.

2. Case study on media councils in Europe

Social media councils could be modelled on press councils, as they enable industry-wide complaint mechanisms and appropriate remedies. Press councils – although mainly active in a professional environment – may provide a helpful starting point for social media councils as they have a long history of dealing with the various issues at stake, such as independence, impartiality, fairness, and effectiveness (including remedies).

First, press and media councils in Europe are all national and apply national codes of ethics, with no pan-European press council. The Alliance of Independent Press Councils of Europe (a network of press and broadcast councils from across Europe) has two "core beliefs" on jurisdiction. First, regulation should be based on nations' differing cultures. Second, a universal code of ethics is impossible, and there should be no supranational codes and regulatory organisations, whether European or global.[60] While these principles concern media ethics codes, such views may also apply online. One criticism of the current content moderation system is that social media companies have "insufficient understanding of linguistic and cultural nuance," an "absence of rigorous human evaluation of context," and problems with "addressing context, widespread variation of language cues and meaning and linguistic and cultural particularities."[61] National content rules and national social media councils may be more appropriate. The "national" scope of the press and media councils doesn't exclude the fact that many of the involved companies are international.

Second, independence is a key principle for press and media councils. For example, the Netherlands Press Council is composed of a chairperson, four (vice) chairs, 10 member-journalists and 10 non-journalist members. The chair is a high-profile journalist, and the four vice-chairs are (former) members of the judiciary. The Netherlands Press Council is established by a foundation called the Stichting Raad voor de Journalistiek, which is composed of the Netherlands Union of Journalists, the Society of Chief-Editors, printed press organisations, and public and commercial broadcasting organisations. This foundation appoints the members of the press council. Another example is Germany, where only publishers and journalists sit on the Press Council board, with no independent representatives.[62] If these models were applied to social media councils, the council would include social media company officials, and be established by a foundation comprising social media interest groups. It is difficult to see how such a social media council would satisfy the principle of independence.

Third, there is not a uniform view on who can complain to press councils. In countries such as Sweden, Denmark, and Ireland, only a person affected by the material can complain. In contrast, in Finland and Germany, the Councils will accept a complaint from any complainant, which could include complaints about general issues of misleading reporting or the failure to separate fact from opinion. There are similar issues for online content moderation decisions. Should any user (or civil society organisation) be able to complain that another user's content should be removed, based on the company's own content rules or terms of service? Of course, press and media councils mainly address issues on content which falls under direct editorial responsibility, and the essential difference with social media companies is that content is generated by users.

Finally, European press councils offer ideas about potential remedies. Most press councils only provide remedies in the form of publication of a decision. The Netherlands Press Council has no power to impose sanctions or to order a correction, rectification, or reply. The vast majority of press councils do not have the power to issue fines, or order corrections, apologies or removal of content. Only one press council, the Swedish Press Council, can impose an administrative fee that is imposed on an upheld complaint, and is tiered depending on the circulation of the publication.[63] In addition, press councils can also operate arbitration procedures.[64]

E-Court Case Studies

1. European dispute resolution mechanisms

The EU has adopted specific legislation in relation to small claim court procedures, out-of-court dispute resolution, and online dispute resolution, which can inform our discussion. Some examples concern EU cross-border initiatives and build upon national systems.

a. Alternative Dispute Resolution Directive

First, there is the Alternative Dispute Resolution Directive of 2013.[65] This directive aims to harmonise rules across the EU to ensure that consumers can voluntarily submit complaints against companies to ADR bodies, which offer independent, impartial, transparent, effective, fast and fair alternative dispute resolution procedures. The ADR Directive applies to procedures for the out-of-court resolution of domestic and cross-border disputes concerning contractual obligations stemming from sales or service

contracts between a trader and a consumer. Crucially, the directive includes "quality requirements" for ADR bodies and procedures, including detailed rules on (a) expertise, independence and impartiality; (b) transparency; (c) effectiveness; and (d) fairness.

First, the rules on *expertise, independence and impartiality* include that ADR officials must possess the necessary knowledge and skills in the field of alternative or judicial resolution of consumer disputes, as well as a general understanding of law; are appointed for a term of office of sufficient duration to ensure the independence of their actions, and are not liable to be relieved from their duties without just cause; are not subject to any instructions from either party or their representatives; are remunerated in a way that is not linked to the outcome of the procedure; and must disclose conflict of interests.

Second, the rules on *transparency* include that ADR bodies must publish details of their expertise, impartiality and independence; the types of disputes handled; procedural rules governing the resolution of a dispute; the costs, if any, to be borne by the parties; the average length of the ADR procedure; and the legal effect of the outcome of the ADR procedure. Further, ADR bodies must also publish annual reports, including the number of disputes received, average time taken to resolve disputes, and rate of compliance, if known.

Third, the rules on *effectiveness* include that the ADR procedures must be available and easily accessible online and offline to both parties; parties must have access to the procedure without being obliged to retain a lawyer or a legal advisor; ADR procedures must be free of charge or available at a nominal fee for consumers; and the outcome of the ADR procedure must be made available within a period of 90 days.

Finally, the rules on *fairness* include that parties have the possibility of expressing their point of view; may seek independent advice or be represented or assisted by a third party at any stage of the procedure; parties are notified of the outcome of the ADR procedure in writing, and are given a statement of the grounds on which the outcome is based. Importantly, parties must, before agreeing to or following a proposed solution, be informed of the legal effect of such a proposed solution, and before expressing their consent to a proposed solution or amicable agreement, be allowed a reasonable period of time to reflect.

Importantly, the directive does not apply to dispute resolution mechanisms where the ADR officials are "employed or remunerated exclusively by an individual trader," but the directive does include a clause that member states can bring such procedures under the directive, where certain requirements are satisfied, including the "specific requirements of independence and transparency." Further, the directive does not apply to consumer complaint-handling systems operated by traders, and attempts made by a judge to settle a dispute in the course of a judicial proceeding concerning that dispute.

Of particular note, the directive contains rules on the binding and non-binding natures of ADR agreements. An agreement between a consumer and a trader to submit complaints to an ADR entity is not binding on the consumer if it was concluded before the dispute has materialised, and if it deprives the consumer of their right to bring an action before the courts to settle the dispute. Further, the solution imposed may be binding on the parties only if they were informed of its binding nature

in advance and specifically accepted this. As such, any proposed e-court model that included an ADR mechanism would need to be consistent with the ADR Directive's rules.

b. European Online Dispute Resolution Regulation

The EU has also enacted the Online Dispute Resolution Regulation.[66] This Regulation established a European Online Dispute Resolution platform, which would facilitate independent, impartial, transparent, effective, fast and fair out-of-court resolution of disputes between consumers and traders online. The Regulation required the European Commission to establish such a platform, designed to be a "single point of entry for consumers and traders seeking the out-of-court resolution" of certain disputes, and free of charge. It applies to out-of-court resolution of disputes concerning contractual obligations stemming from online sales or service contracts. In 2016, the Commission established the European Online Dispute Resolution platform. In 2017, over 24,000 consumer complaints were lodged with the platform.[67] Over a third of the complaints concerned cross-border purchases within the EU; and most complaints concerned clothing and footwear, airline tickets and information and communication technology goods.

The ODR Regulation sets down the rules and procedure for the operation of the platform. First, a consumer can submit an electronic complaint form; and the ODR platform automatically notifies the company, which has a period of 10 days to reply, of the request. Second, the ODR platform facilitates the exchange of messages directly through a dashboard, and permits sending attachments such as product photos, and scheduling an online meeting. Third, parties have 30 days from the submission of the complaint to agree on a dispute resolution body; and the ODR platform will provide recommended dispute resolution bodies. However, the parties may choose outside of this recommended list. Once the parties have agreed on a dispute resolution body, the ODR platform will provide the ODR body with the consumer's complaint. The body then has three weeks to inform the parties whether it accepts jurisdiction to handle the complaint. Once the dispute resolution body has accepted jurisdiction, it will handle the complaint in line with its procedures, and must offer a suggested solution with 90 days. The final decision of the body will be available on the ODR platform. Whether or not the suggested solution is legally binding on the parties depends upon the rules of the particular dispute resolution body handling the complaint.

The European Online Dispute Resolution platform is an example of an online dispute resolution mechanism established by legislation, and facilitating online negotiation, and independent dispute resolutions. There is a network of national online dispute resolution bodies as part of the platform, and EU member states are required to designate at least one such body. A European social media dispute resolution platform could be created along these lines to facilitate dispute resolution between users and social media companies.

c. European Small Claims Procedure

In addition to online alternative dispute resolution mechanisms, the EU has enacted a European Small Claims Procedure Regulation,[68] which was designed to simplify and speed up litigation concerning small claims in cross-border cases, and to reduce costs. It was designed with a number of features: (a) an online procedure that does not require a physical presence in court; (b) eliminates the intermediate proceedings necessary to enable recognition and enforcement of judgments from another member

state; (c) is an alternative to national procedures; and (d) a decision given in the European Small Claims Procedure is recognised and enforceable in all member states. The regulation applies to cross-border cases involving civil and commercial matters, where the value of a claim does not exceed 5,000 euro. Notably, the regulation does not apply to "violations of privacy and of rights relating to personality, including defamation."

The European Small Claims Procedure is now available on the EU's European e-Justice online platform,[69] and operates as follows. To start the procedure, a claim form is sent, with relevant supporting documents, such as receipts, invoices, etc., to the national court with jurisdiction. Member states must ensure the procedure is accessible through relevant national websites. Once the court receives the application, it completes a certain form, and within 14 days the court should serve a copy of it, along with the form, on the defendant. The defendant has 30 days to reply, by filing another form. The court must then send a copy of any reply to the plaintiff within 14 days.

Within 30 days of receiving the defendant's answer, the court must either give a judgment on the small claim, or summon the parties to an oral hearing. The Regulation provides that this oral hearing should use any appropriate distance communication technology, such as videoconference or teleconference, unless it would be unfair. Further, a party summoned to be physically present at an oral hearing may request the use of distance communication technology, and it is not necessary to be represented by a lawyer. The court is required to use the "simplest and least burdensome method of taking evidence" and the court "shall not require the parties to make any legal assessment of the claim." Whenever appropriate, the court must seek to reach a settlement between the parties. Finally, once a decision is reached, a certificate is issued by the court, and the judgment is enforceable in all member states, without any further formalities.

The European Small Claims Procedure is an e-court model based on online judicial adjudication, and established by legislation. Social media disputes could follow a similar online procedure, with online judicial decision-making, and the imposition of time limits.

d. Business-to-Platform Regulation

It may also be helpful to refer to EU legislation applicable to disputes between certain online platforms and their *business* users, called the Business-to-Platform Regulation 2019.[70] While the regulation is only applicable to platforms offering online marketplaces (e.g., Amazon, eBay, and mobile app stores) it is still relevant to some questions on content moderation decisions.

First, it has rules on platform *terms and conditions*, including that the rules must clearly set out the grounds for decisions to suspend, terminate, or restrict a business user's account. Platforms must give business users notice of any planned changes to the terms and conditions. Second, there are detailed rules on *account suspension and terminations*, including the requirement that when a platform decides to suspend or terminate a business user account, it must provide a statement of reasons for that decision. Third, platforms must establish *internal complaint-handling systems*, including the obligation to provide individualised decisions drafted in plain and intelligible language. Fourth, and importantly, the regulation requires platforms to identify *impartial and independent mediators* in their terms and conditions which they are willing to engage for out-of-court dispute resolution. Article 12 sets downs rules for these mediators, including that they must be impartial and independent, affordable for business users,

easily accessible, and have sufficient understanding of general business-to-business commercial relations, enabling them to contribute effectively to dispute settlement.

The Business-to-Platform Regulation is the first piece of sector-specific regulation for online platforms. While only applicable to business users, it attempts to bring fairness and transparency for these business users. The regulation's rules on clear terms and conditions, internal complaint-handling, and account suspension could serve as a helpful model for best practices for social media companies. The regulation's rules on independent dispute resolution could also be used for best practices. For example, social media companies could recognise third-party online dispute resolution bodies they would be willing to engage to resolve content moderation disputes.

2. Case study on e-courts in the Netherlands

The term e-court covers a broad spectrum of online dispute resolution in the Netherlands, including private and public initiatives. As regards public initiatives, e-courts are often a way of digital litigation in national courts, for example the former "eKantonrechter."[71] Private initiatives fall under alternative dispute resolution and are usually a form of arbitration.[72] The best-known private Dutch e-courts are Stichting e-Court,[73] Stichting DigiTrage[74] and Stichting Arbitrage Rechtspraak Nederland.[75]

E-courts began in the Netherlands in 2009 with the Stichting e-Court.[76] It was followed by Stichting Arbitrage Rechtspraak Nederland in 2012,[77] the eKantonrechter in 2013[78] and Stichting DigiTrage in 2014.[79] All these initiatives had the goal to lower the threshold for dispute resolution by introducing a simplified, faster and (as for private initiatives) less costly procedure. However, only the e-courts of Stichting DigiTrage and Stichting Arbitrage Rechtspraak Nederland are still in use. The eKantonrechter was terminated in 2018 by the Dutch Council for the Judiciary because it "would no longer fit well within the ambition to simplify."[80] Also, Stichting e-Court has been stalled since the beginning of 2018 due to negative reporting[81] and related lawsuits.[82]

The Dutch Code of Civil Procedure provides a national legal basis for e-courts' provisions on arbitration.[83] In these cases, the arbitration institutions will be often included in the general terms and conditions as agreed between parties.[84] An opponent can opt out within (at least) one month.[85] To enforce the decision, an instrument permitting the enforcement issued by a national court (an exequatur) is still required.[86] The national court examines the arbitration decision summarily.[87] No regulatory body is involved. The Netherlands Arbitration Institute[88] promotes alternative dispute resolution by, inter alia, publishing draft texts for arbitration clauses.[89] The eKantonrechter programme operated by courts also found its legal basis in the Dutch Code of Civil Procedure.[90]

Online dispute resolution initiatives in the Netherlands differ in the types of disputes that they address. Stichting e-Court handles mainly debt collection cases of the largest Dutch insurance companies (more than 20,000 cases in 2017).[91] Five out of six decisions published by the court concern unpaid contributions for health insurance.[92] Stichting DigiTrage focusses on debt collection cases of small and medium-size enterprises.[93] The eKantonrechter was allowed to handle essentially (uncomplicated) civil matters up to 25,000 euro (approximately 27,800 USD) and labour and rental matters,[94] but has only handled fourteen cases in a four-year period.[95] One case involved claiming damages for "non-conformity" in the sale of a house.[96] It is unclear which specific type of cases Stichting Arbitrage Rechtspraak Nederland deals with.

The general procedure at e-courts is as follows. First, the party seeking redress submits an application on an online platform. Then, the defendant lodges a statement of defence on the same platform. A round of online adversarial hearings occurs. If necessary, an oral hearing can be held (online or offline), and witnesses summoned. Finally, judgment will be rendered. Each body can deviate from this procedure.[97] Occasionally, appeal against the arbitration decision is possible.[98]

Compared with regular courts, e-courts generally offer shorter deadlines and faster procedures, lower and clearer costs, more convenience (litigation from home and the possibility to access your dossier anytime and anywhere) and the possibility to litigate in comprehensible language. The main difficulty of (private) e-courts is that they lack transparency. Decisions are rarely published, and a list of arbitrators is not always available.[99] Another major concern is that e-courts financially depend on large parties which often use this type of dispute resolution, such as large insurance companies. This can influence negatively the equality of arms, which requires a fair balance between parties.[100] Another difficulty especially relating to Stichting e-Court is that arbitrators rule in absentia in favour of the applicant.[101] Besides, use of the term "e-court" for arbitration purposes can be misleading.

Multiple questions arise about Dutch e-court initiatives, particularly related to the essential criteria mentioned above, and the requirements of the ADR Directive. First, all the e-courts claim to be independent. However, it is not clear whether there is an arbitrators' term of office and whether this term is sufficiently long to guarantee their independence, which raises questions under the ADR Directive's rules on expertise, independence, impartiality, and transparency. Second, there can be an appearance of partiality because the arbitration institution depends financially, albeit indirectly, on professional parties that include the institution in their general terms and conditions.[102] Besides, a lot of arbitrators are attorneys. It can be questioned whether they have (without specific training) the appropriate skills to judge impartially. This again raises issues under the rules on expertise, independence and impartiality in the ADR Directive. Third, many e-court initiatives lack transparency as decisions are never or rarely published, and a list of arbitrators is not always available. None of the websites contain annual reports, which is required under transparency rules.

Fourth, e-courts seem effective, as they are simple, fast and relatively cheap to use, and assistance by a lawyer is not mandatory. However, not all procedures are also available offline, which is required by the rules on effectiveness. Finally, professional parties use an arbitration institution more often than a consumer. This can have a negative effect on the equality of arms. Apart from that, there are concerns about whether consumers are properly informed that the arbitration institution is not a regular court. This could constitute an infringement of the rules on fairness.

3. Case study on e-courts in North America

There are some excellent examples in North America of e-court initiatives that involve courts adopting online case resolution systems, where individuals with minor disputes engage through an online platform with police, prosecutors, and judges. First is Michigan's Online Case Review programmes for resolving traffic disputes, warrant disputes and misdemeanours.[103] It comprises online platforms for defendants to submit their cases, including all arguments or explanations, such as why they cannot pay their fines. It allows police and prosecutors to review cases before a judge makes a decision. The online format provides for the resolution of traffic disputes without in-person court appearances.

There are further incentives for individuals, under which resolutions may be reached with prosecutors, resulting in, for example, no driver licence points.

Second, some e-court initiatives go beyond the digitisation of court systems (electronic filings, online hearings, etc.), and involve online negotiation, mediation, arbitration, and adjudication. One leading example is the Civil Resolution Tribunal in British Columbia, which is "Canada's first online tribunal," and was established under British Columbia's Civil Resolution Tribunal Act 2012.[104] The purpose of the Civil Resolution Tribunal is to encourage early resolution of disputes through online dispute resolution, while preserving adjudication as a last resort. Under the act, the tribunal has authority to resolve certain small claims, up to a value of CAD$5,000 (it has no jurisdiction over claims involving libel); and also provides that certain small claims *must* go through the tribunal before going to provincial court.

The tribunal process follows a stepped online-dispute resolution process. First, a free, anonymous, and confidential online platform helps complainants assess their problem and decide the best option for how to proceed. Second, if the user cannot resolve the issue through the Solution Explorer platform, the process moves to an online dispute resolution portal, which begins with party-to-party negotiation. It is an online guided and structured negotiation. While it is an independent process, tribunal staff monitor the negotiations and are available to provide case-specific suggestions and support. Third, if there no is agreement, the process moves to the facilitated settlement phase. This involves neutral tribunal case managers who are dispute resolution experts and discuss the issues and settlement options. Under the act, if the parties reach a facilitated agreement, they can apply to the tribunal for a consent resolution order. If parties still cannot reach a solution, the claim proceeds to resolution by tribunal hearing. An independent tribunal member will make a decision about the dispute. If hearings are not needed, the tribunal member may render a decision based solely on digital evidence and submissions. The Act has detailed rules on enforcement of tribunal orders, and judicial review of these decisions.

Although not currently operational, it is worth mentioning a similar initiative, but involving members of the judiciary, rather than tribunal officials. The UK Civil Justice Council's advisory group has recommended the creation of an e-court system (HM Online Courts), designed to reduce the need for judges in many cases.[105] There would be three tiers: the first tier would be online evaluation, to help users with grievances evaluate their problems, and understand both their entitlements and the options available to them. This would be a form of information and diagnostic service and would be available at no cost to court users. The second tier would be online facilitation, which would involve trained and experienced facilitators, working online, who can review papers and statements from parties, and then help by mediating, advising, or encouraging them to negotiate. This stage would be designed to bring many if not most disputes to speedy, fair conclusions without the involvement of judges; and users would incur a court fee. Then tier three would be online judges, where cases would be decided on an online basis, largely on the basis of papers submitted to them electronically, as part of a structured but still adversarial system of online pleading and argument. The decisions of online judges would be binding and enforceable, enjoying the same status as decisions made by judges in traditional courtrooms. A court fee would be payable, but much lower than normal court fee.

Notes

[1] Dr. Heidi Tworek is assistant professor at the University of British Columbia as well as a non-resident fellow at the German Marshall Fund of the United States and the Canadian Global Affairs Institute. Her book *News from German: The Competition to Control World Communications, 1900-1945*, was published in 2019 by Harvard University Press.

[2] Dr. Ronan Ó Fathaigh is a researcher at the Institute for Information Law (IViR), University of Amsterdam, whose doctoral dissertation at Ghent University was on freedom of expression and the chilling effect.

[3] Lisanne Bruggeman participated in the Glushko & Samuelson Information Law and Policy Lab at the Institute for Information Law (IViR), University of Amsterdam, and is completing the master's programme in information law at the University of Amsterdam.

[4] Dr. Chris Tenove is a postdoctoral research fellow at the University of British Columbia who studies political theory and international relations, with a focus on issues of democracy, public policy, global governance, and digital politics.

[5] Josh Constine, "Instagram will let you appeal post takedowns," *TechCrunch*, 9 May 2019, https://techcrunch.com/2019/05/09/instagram-vaccine-hashtags/

[6] Jillian C. York and Corynne McSherry, "Content Moderation is Broken. Let Us Count the Ways," Electronic Frontier Foundation, 29 April 2019, https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways.

[7] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, A/HRC/38/35, 6 April 2018, para. 38, https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35.

[8] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, A/HRC/38/35, 6 April 2018, para. 72.

[9] On platform governance, see the series of essays, "Models for Platform Governance," Centre for International Governance Innovation, 29 October 2019, https://www.cigionline.org/publications/models-platform-governance.

[10] Facebook, Community Standards Enforcement Report, May 2019 - Hate Speech - How much content did we restore after removing it?, https://transparency.facebook.com/community-standards-enforcement#hate-speech.

[11] Jonathan Vanian, "Google's Hate Speech Detection A.I. Has a Racial Bias Problem," *Fortune*, 16 August 2019,

https://fortune.com/2019/08/16/google-jigsaw-perspective-racial-bias/ .

[12] For empirical analysis of the chilling effect, see Jonathon W. Penney, "Internet surveillance, regulation, and chilling effects online: a comparative case study" (2017) 6(2) *Internet Policy Review*, DOI:10.14763/2017.2.692. For a European discussion, see Ronan Ó Fathaigh, *Article 10 and the Chilling Effect: A critical examination of how the European Court of Human Rights seeks to protect freedom of expression from the chilling effect* (PhD Thesis, Ghent University, 2019), https://biblio.ugent.be/publication/8620369.

[13] See *Dink v. Turkey* (App. nos. 2668/07, 6102/08, 30079/08, 7072/09 and 7124/09) 14 September 2010, para. 137; Human Rights Committee, General comment No. 34, CCPR/C/GC/34, 12 September 2011, para. 7; and Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, A/HRC/38/35, 6 April 2018, para. 6. See Brittan Heller and Joris van Hoboken, "Freedom of Expression: A Comparative Summary of United States and European Law," working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (2019), https://www.ivir.nl/publicaties/download/TWG_Freedom_of_Expression.pdf.

[14] As stated in the memo presented by European Court of Human Rights Judge Róbert Spanó during the Bellagio meeting in November 2019. Other parts of the memo are reflected in the paragraph and elsewhere in the paper.

[15] Heidi Tworek, "How Transparency Reporting Could Incentivize Irresponsible Content Moderation," Centre for International Governance Innovation, 10 December 2019, https://www.cigionline.org/articles/how-transparency-reporting-could-incentivize-irresponsible-content-moderation.

[16] Amy J. Schmitz, "Expanding Access to Remedies through E-Court Initiatives" (2019) 67 *Buffalo Law Review* 89.

[17] More input on the criteria of independence can be derived from: K. Irion, W. Schulz, W., P. Valcke, *The Independence of the Media and Its Regulatory Agencies: Shedding New Light on Formal and Actual Independence Against the National Context*, Intellect, Bristol UK / Chicago USA, 2014.

[18] We're not discussing the definition of self-regulation (or co-regulation) as such. There is rich literature on this: e.g., Michael Latzer, Natascha Just, & Florian Saurwein, "Self- and co-regulation: Evidence, legitimacy and governance choice," in Monroe E. Price and Stefaan Verhulst, eds., *Routledge Handbook of Media Law*, New York, 2013, 373-398.

[19] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, A/HRC/38/35, 6 April 2018, para. 25.

[20] The Twitter Trust and Safety Council, https://about.twitter.com/en_us/safety/safety-partners.html#Trust-council.

[21] Jane Wakefield, "Google's ethics board shut down," *BBC News*, 5 April 2019, https://www.bbc.com/news/technology-47825833.

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

[23] YouTube Trusted Flagger program, https://support.google.com/youtube/answer/7554338?hl=en.

[24] See Wikipedia: Arbitration Committee, https://en.wikipedia.org/wiki/Wikipedia:Arbitration_Committee. See also Corinne Ramey, "The 15 People Who Keep Wikipedia's Editors From Killing Each Other," *The Wall Street Journal*, 7 May 2018, https://www.wsj.com/articles/when-wikipedias-bickering-editors-go-to-war-its-supreme-court-steps-in-1525708429; and Piotr Konieczny, "Decision making in the self-evolved collegiate court: Wikipedia's Arbitration Committee and its implications for self-governance and judiciary in cyberspace," (2017) 32(6) *International Sociology* 755.

[25] Global Network Initiative, Our Mission, https://globalnetworkinitiative.org/team/our-mission.

[26] Internet Watched Foundation, https://www.iwf.org.uk/become-a-member/join-us/our-members.

[27] See, for example, Kristina Irion et al., *The independence of media regulatory authorities in Europe* (European Audiovisual Observatory, 2019), https://rm.coe.int/the-independence-of-media-regulatory-authorities-in-europe/168097e504.

[28] See Andrei Richter, *Disinformation in the media under Russian law* (European Audiovisual Observatory, 2019), p. 12, https://rm.coe.int/disinformation-in-the-media-under-russian-law/1680967369.

[29] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, A/HRC/38/35, 6 April 2018, para. 72.

[30] ARTICLE 19, *Self-regulation and 'hate speech' on social media platforms* (2018), https://www.article19.org/resources/self-regulation-hate-speech-social-media-platforms/.

[31] Chris Tenove, Heidi Tworek, and Fenwick McKelvey, *Poisoning Democracy: How Canada Can Address Harmful Speech Online* (2018), https://ppforum.ca/publications/poisoning-democracy-what-can-be-done-about-harmful-speech-online/.

[32] Mark Zuckerberg, "A Blueprint for Content Governance and Enforcement," Facebook, 15 November 2018, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/?hc_location=ufi.

[33] Stanford Global Digital Policy Incubator, Article 19, and UN Special Rapporteur on Freedom of Opinion and Expression, *Social Media Councils: From Concept to Reality - Conference Report* (2019), https://fsi.stanford.edu/content/social-media-councils-concept-reality-conference-report.

[34] ARTICLE 19, *The Social Media Councils: Consultation Paper* (Article 19, 2019), https://www.article19.org/resources/social-media-councils-consultation/.

[35] ARTICLE 19, *The Social Media Councils: Consultation* Paper (2019), https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf.

[36] Heidi Tworek, *Social Media Councils*, Centre for International Governance Innovation, October 2019, https://www.cigionline.org/articles/social-media-councils.

[37] Jacinda Ardern, Prime Minister of New Zealand, "Significant progress made on eliminating terrorist content online," 24 September 2019, https://www.beehive.govt.nz/release/significant-progress-made-eliminating-terrorist-content-online.

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

[39] Facebook, Oversight Board Charter, September 2019, p. 4, https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf.

[40] Ibid.

[41] Ibid., p. 7.

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

[43] For further suggestions, see Heidi Tworek, "Social Media Councils," Centre for International Governance Innovation, 2019, https://www.cigionline.org/articles/social-media-councils.

[44] Canadian Radio-television and Telecommunications Commission, *Harnessing Change: The Future of Programming Distribution in Canada*, 2018, https://crtc.gc.ca/eng/publications/s15/.

[45] Mandhane, Renu, and Marie-Claude Landry, "Addressing Discriminatory Advertising on Facebook in Canada," Ontario Human Rights Commission, 7 June 2019, http://www.ohrc.on.ca/en/news_centre/addressing-discriminatory-advertising-facebook-canada.

[46] Amy J. Schmitz, "Expanding Access to Remedies through E-Court Initiatives" (2019) 67 *Buffalo Law Review* 89, 92.

[47] Civil Justice Council's Online Dispute Resolution Advisory Group, *Online Dispute Resolution for Low Value Civil* Claims (2015), https://www.judiciary.uk/wp-content/uploads/2015/02/Online-Dispute-Resolution-Final-Web-Version.pdf.

[48] Schmitz, 100.

[49] The Office of Communications Annual Report & Accounts For the period 1 April 2018 to 31 March 2019, p. 149, https://www.ofcom.org.uk/__data/assets/pdf_file/0024/156156/annual-report-18-19.pdf.

[50] Standing Committee on Access to Information, Privacy and Ethics, *Democracy Under Threat: Risks and Solutions in the Era of Disinformation and Data Monopoly,* (Committee Report 17. Ottawa, ON: House of Commons, Canada, 2018) http://www.ourcommons.ca/DocumentViewer/en/42-1/ETHI/report-17. See also (Government of Canada 2019). While some groups welcomed this, others have expressed concern about this vague aspiration (see, for instance, Zwibel 2019).

[51] Liberal Party of Canada, "Forward: A Real Plan for the Middle Class," Federal Liberal Agency of Canada, 2019, p. 47, https://2019.liberal.ca/wp-content/uploads/sites/292/2019/09/Forward-A-real-plan-for-the-middle-class.pdf.

[52] E.g. Mack Lamoureux, "YouTube Pulls Canadian Anti-Islam Vlogger Following Huge Defamation Loss," *Vice*, 14 May 2019, https://www.vice.com/en_ca/article/597ddk/youtube-pulls-canadian-anti-islam-vlogger-kevin-j-johnston-following-record-defamation-lawsuit-loss

[53] Chris Tenove, Chris, Heidi J. S. Tworek, and Fenwick McKelvey, "Poisoning Democracy: How Canada Can Address Harmful Speech Online," 2018, Ottawa, ON: Public Policy Forum, https://www.ppforum.ca/wp-content/uploads/2018/11/PoisoningDemocracy-PPF-1.pdf.

[54] National News Media Council, "Frequently Asked Questions," https://mediacouncil.ca/frequently-asked-questions/.

[55] Canadian Broadcast Standards Council, https://www.cbsc.ca/about-us/.

[56] Canadian Radio-television and Telecommunications Commission, *Harnessing Change: The Future of Programming Distribution in Canada*, 2018, https://crtc.gc.ca/eng/publications/s15/.

[57] Mandhane, Renu, and Marie-Claude Landry, "Addressing Discriminatory Advertising on Facebook in Canada," Ontario Human Rights Commission, 7 June 2019, http://www.ohrc.on.ca/en/news_centre/addressing-discriminatory-advertising-facebook-canada.

[58] Section 13 of the Act barred individuals or groups from using telecommunications – later expanded to include the internet – to distribute messages likely to expose a person or group to hatred or contempt based on their race, ancestry, religion, and other characteristics.

[59] Report of the Standing Committee on Justice and Human Rights, "Taking Action to End Online Hate," June 2019, https://www.ourcommons.ca/Content/Committee/421/JUST/Reports/RP10581008/justrp29/justrp29-e.pdf

[60] Alliance of Independent Press Councils in Europe, http://www.aipce.net/aboutAipce.html.

[61] See Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, A/HRC/38/35, 6 April 2018, para. 54.

[62] See, for example, Lara Fielden, Regulating the Press: A Comparative Study of International Press Councils (Reuters Institute for the Study of Journalism, 2012), https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-11/Regulating%20the%20Press.pdf.

[63] Ibid., p. 15.

[64] See, eg., IMPRESS. Arbitration, https://www.impress.press/regulation/arbitration.html.

[65] Directive 2013/11/EU on alternative dispute resolution for consumer disputes (ADR Directive).

[66] Regulation (EU) No 524/2013 on online dispute resolution (ODR Regulation).

[67] See the European Online Dispute Procedure, https://ec.europa.eu/consumers/odr/main/?event=main.home.howitworks.

[68] Regulation (EC) No 861/2007 of the European Parliament and of the Council of 11 July 2007 establishing a European Small Claims Procedure – consolidated text of 14 June 2017, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02007R0861-20170714.

[69] https://e-justice.europa.eu/content_small_claims-42-en.do.

[70] Regulation (EU) 2019/1150 on promoting fairness and transparency for business users of online intermediation services (Business-to-Platform Regulation).

[71] See, Nu ook online rechtspraak voor burgers bij de eKantonrechter, 3 June 2014, https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Raad-voor-de-rechtspraak/Nieuws/Paginas/Nu-ook-online-rechtspraak-voor-burgers-bij-de-eKantonrechter.aspx; and see also: J.W. Langelaar, 'De eKantonrechter', *Tijdschrift voor de Procespraktijk* 2015-3.

[72] Another type of alternative dispute resolution e-courts in certain cases deal(t) with is "binding advice."

[73] http://www.e-court.nl/; see also: D.E. Thiescheffer, *E-court naast overheidsrechtspraak: Online rechtspraak als alternatieve geschilbeslechting en de garanties voor consumenten* (Celsus juridische uitgeverij, 2018).

[74] Digitage, "Digitale arbitragevoor incassozaken," https://www.digitrage.nl/.

[75] Stichting Arbitrage Rechtspraak Nederland, https://www.arbitragerechtspraak.nl/.

[76] "Nieuwe rechtbank opent haar deuren," 12 January 2010, http://www.e-court.nl/persberichten/.

[77] Trade register of the Dutch Chamber of Commerce (consulted on 26 September 2019).

[78] "Nu ook online rechtspraak voor burgers bij de eKantonrechter," 3 June 2014, https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Raad-voor-de-rechtspraak/Nieuws/Paginas/Nu-ook-online-rechtspraak-voor-burgers-bij-de-eKantonrechter.aspx

[79] Emerce, "Digitrage gestart met online geschillenbeslechting," 2 February 2014, https://www.emerce.nl/wire/digitrage-gestart-online-geschillenbeslechting

[80] "Reset digitalisering van de Rechtspraak," 10 April 2018, https://www.rechtspraak.nl/SiteCollectionDocuments/2018-brief-reset-digitalisering.pdf, p. 5.

[81] See, Karlijn Kuijpers, Thomas Muntz and Tim Staal, "Vonnis te koop," De Groene Amsterdammer, 17 January 2018, https://www.groene.nl/artikel/vonnis-te-koop; Sociaal Werk Nederland, "Rechtspraak op bestelling?! Stop commerciële rechtspraak," January 2018, https://schuldinfo.nl/fileadmin/Publicaties/Rechtspraak_op_bestelling_Stop_commerciele_rechtspraak.pdf; and the responses of Stichting e-Court: http://www.e-court.nl/wp-content/uploads/2018/01/Persbericht-2018-01-19-reactie-publiciteit.pdf; http://www.e-court.nl/wp-content/uploads/2018/01/Reactie-LOSR-2018-01-22.pdf.

[82] See: e-Court, e-Court naar het Europese Hof van Justitie, 16 February 2018, http://www.e-court.nl/wp-content/uploads/2018/02/Persbericht-2018-02-16-e-Court-naar-EHvJ.pdf; and e-Court, "Dutch 'e-Court' takes

Supreme Court justices and other magistrates to Court," 16 November 2018, http://www.e-court.nl/wp-content/uploads/2018/11/Persbericht-2018-11-16-RvdR_Eng.pdf.

[83] Articles 1020-1076 Wetboek van Burgerlijke Rechtsvordering.

[84] See also: Article 1020 Wetboek van Burgerlijke Rechtsvordering.

[85] Article 6:236, paragraph n, Burgerlijk Wetboek.

[86] Article 1062, paragraph 1, Wetboek van Burgerlijke Rechtsvordering.

[87] Article 1063, paragraph 1, Wetboek van Burgerlijke Rechtsvordering.

[88] Netherlands Arbitration Institute, https://www.nai-nl.org/en/ .

[89] Netherlands Arbitration Institute, "Clauses," https://www.nai-nl.org/en/documents/clauses/ .

[90] Article 96 Wetboek van Burgerlijke Rechtsvordering.

[91] "Robotrechter e-Court reageert: 'De rechterlijke macht werkt ons tegen'," *Het Financieele Dagblad*, 28 January 2018, https://fd.nl/economie-politiek/1239049/robotrechter-e-court-reageert-de-rechterlijke-macht-werkt-ons-tegen.

[92] See decision numbers 21-04-2017 IPVRW 24009030000, 30-06-2017 IPVRW 24363052000, 17-11-2017 IPVRW 24562467000, 17-11-2017 IPVRW 24590256000 and 29-09-2017 IPVRW 24127585000; available here: http://www.e-court.nl/uitspraken.

[93] Digitrage, "Geschil Aanmelden," https://www.digitrage.nl/geschil-aanmelden.html; and Digitrage, "Bedrijven met Digitrage-beding," https://www.digitrage.nl/bedrijven-met-digitrage-beding.html.

[94] Article 96, paragraph 1, and article 93 Wetboek van Burgerlijke Rechtsvordering.

[95] "Samen naar de eKantonrechter kan niet meer," *Mr.*, 29 May 2018, https://www.mr-online.nl/samen-naar-de-ekantonrechter-kan-niet-meer/.

[96] ECLI:NL:RBOBR:2014:1145, 13 March 2014, https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBOBR:2014:1145.

[97] For more information, see the procedural regulations on the institutions' websites.

[98] Article 30 Procesreglement DigiTrage 01-01-2015; Article 16 Procesreglement e-Court 2017; Article 30 Arbitragereglement Stichting Arbitrage Rechtspraak Nederland 15-08-2018.

[99] Only six decisions have been published by Stichting e-Court: http://www.e-court.nl/uitspraken/. Only Stichting DigiTrage (https://www.digitrage.nl/over-digitrage/de-organisatie.html) and Stichting e-Court (http://www.e-court.nl/persberichten/) have published lists of arbitrators on their websites.

[100] See also: D.E. Thiescheffer, E-court naast overheidsrechtspraak: Online rechtspraak als alternatieve geschilbeslechting en de garanties voor consumenten' (Celsus juridische uitgeverij, 2018), p. 81.

[101] Article 12, paragraph 2, Procesreglement e-Court 2017, unless the arbitrator judges the claim to be unlawful or unfounded.

[102] See also: C.N.J. de Vey Mestdagh and A. Kamphorst, "e-Court of de moderne Prometheus vs. het recht, een achterhoedegevecht?," *Tijdschrift voor Internetrecht* (2018), p. 14.

[103] See, for example, 74th District Court Online Case Review, https://www.baycounty-mi.gov/News/74th-District-Court-Online-Case-Review.aspx.

[104] See Civil Resolution Tribunal, https://civilresolutionbc.ca; and the Civil Resolution Tribunal Act 2012, http://www.bclaws.ca/civix/document/id/complete/statreg/12025_01.

[105] Civil Justice Council's Online Dispute Resolution Advisory Group, Online Dispute Resolution for Low Value Civil Claims (2015), p. 3.