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.