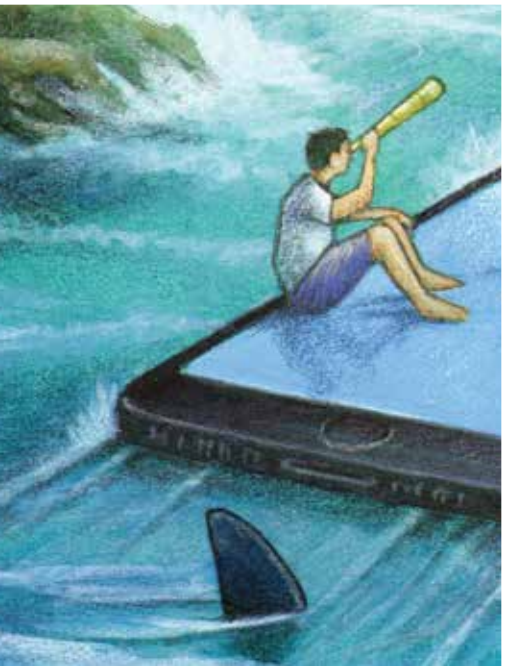
One in a Series of Working Papers from the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression



Co-Chairs Reports of the Three Sessions of the Transatlantic Working Group

The Ditchley Park Session

(Feb. 27-March 3, 2019, at Ditchley Park, UK)
Susan Ness and Nico van Eijk

The Santa Monica Session

(May 9-12, 2019, at the Annenberg Community Beach House, Santa Monica, California, U.S.)
Susan Ness and Nico van Eijk

The Bellagio Session

(Nov. 13-16, 2019, at the Rockefeller Foundation Center, Bellagio, Italy)
Susan Ness and Marietje Schaake

TRANSATLANTIC WORKING GROUP

The Transatlantic Working Group Papers Series

Co-Chairs Reports

Co-Chairs Reports from TWG's Three Sessions: Ditchley Park, Santa Monica, and Bellagio.

Freedom of Expression and Intermediary Liability

Freedom of Expression: A Comparative Summary of United States and European Law
B. Heller & J. van Hoboken, May 3, 2019.

Design Principles for Intermediary Liability Laws
J. van Hoboken & D. Keller, October 8, 2019.

Existing Legislative Initiatives

An Analysis of Germany's NetzDG Law
H. Tworek & P. Leerssen, April 15, 2019.

The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications
J. van Hoboken, May 3, 2019.

Combating Terrorist-Related Content Through AI and Information Sharing
B. Heller, April 26, 2019.

The European Commission's Code of Conduct for Countering Illegal Hate Speech Online: An Analysis of Freedom of Expression Implications
B. Bukovská, May 7, 2019.

The EU Code of Practice on Disinformation: The Difficulty of Regulating a Nebulous Problem
P.H. Chase, August 29, 2019.

A Cycle of Censorship: The UK White Paper on Online Harms and the Dangers of Regulating Disinformation
P. Pomerantsev, October 1, 2019.

U.S. Initiatives to Counter Harmful Speech and Disinformation on Social Media
A. Shahbaz, June 11, 2019.

ABC Framework to Address Disinformation

Actors, Behaviors, Content: A Disinformation ABC: Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses
C. François, September 20, 2019.

Transparency and Accountability Solutions

Transparency Requirements for Digital Social Media Platforms: Recommendations for Policy Makers and Industry
M. MacCarthy, February 12, 2020.

Dispute Resolution and Content Moderation: Fair, Accountable, Independent, Transparent, and Effective
H. Tworek, R. Ó Fathaigh, L. Bruggeman & C. Tenove, January 14, 2020.

Algorithms and Artificial Intelligence

An Examination of the Algorithmic Accountability Act of 2019
M. MacCarthy, October 24, 2019.

Artificial Intelligence, Content Moderation, and Freedom of Expression
E. Llansó, J. van Hoboken, P. Leerssen & J. Harambam, February 26, 2020.

www.annenbergpublicpolicycenter.org/twg

A project of the Annenberg Public Policy Center of the University of Pennsylvania in partnership with
The Annenberg Foundation Trust at Sunnylands and the Institute for Information Law of the University of Amsterdam,
with support from the Embassy of the Kingdom of the Netherlands in Washington, D.C.

Co-Chairs Reports of the Three Sessions of the Transatlantic Working Group

The Ditchley Park Session

(Feb. 27-March 3, 2019, at Ditchley Park, UK)

—*Susan Ness and Nico van Eijk*

The Santa Monica Session

(May 9-12, 2019, at the Annenberg Community Beach House, Santa Monica, California, U.S.)

—*Susan Ness and Nico van Eijk*

The Bellagio Session

(Nov. 13-16, 2019, at the Rockefeller Foundation Center, Bellagio, Italy)

—*Susan Ness and Marietje Schaake*

Co-Chairs Report No. 1: The Ditchley Park Session

Susan Ness, Annenberg Public Policy Center
Nico van Eijk, Institute for Information Law, University of Amsterdam May
2, 2019

Introduction & mission statement

The Transatlantic High Level Working Group on Content Moderation and Freedom of Expression (TWG) held its inaugural meeting at Ditchley Park in the United Kingdom from February 27 to March 3, 2019. Comprised of leading academics, policy makers, and industry representatives, the Working Group convened to discuss the future of freedom of expression in the digital age. This report offers an overview of key outcomes.

Freedom of expression is one of the cornerstones of democracy and international human rights law. Yet this right has never been absolute: democratic societies have deemed certain types of speech so harmful that they are unacceptable. Historically, hate speech and incitement to violence frequently have been subject to restrictions. Deception and false representation have also been found unworthy of protection under certain circumstances.

These types of harmful speech are as old as history, but the internet allows them to propagate at unprecedented speed and scale. Politicians, policy makers, the tech community, and citizens on both sides of the Atlantic are grappling with these new phenomena, considering and often adopting initiatives to restrict "unwanted" content online. Despite best intentions, such efforts have the potential to restrict rightful freedom of expression. The momentum to regulate the (perceived) threats of hate speech and viral deception therefore risks undermining the very democratic systems governments and politicians seek to protect. And despite the internet's transnational reach most measures are considered in national contexts, notwithstanding potential global effects.

The Transatlantic Working Group was formed in response to these trends to develop concrete tools, guidelines, and recommendations to help policy makers navigate the challenges of governing content in the digital age.

Our discussion took into account the many platforms that foster this global conversation – not just the large social media companies and search engines, such as Facebook and Google, which are often the focus of initiatives to address unwanted content, but also smaller European, American and, indeed, global platforms as well as nonprofit, crowd-sourced informational services.

This session of the Transatlantic Working Group explored in depth hate speech and violent extremist content. To this end, we examined four different initiatives designed to address hate speech and incitement to terrorism:

- Germany's Network Enforcement Law (NetzDG)
- The European Union's proposed Terrorism Content Regulation
- The European Union's (voluntary) Code of Conduct for Countering Illegal Hate Speech Online
- The Global Internet Forum to Counter Terrorism's Hash-Sharing Database

Briefing papers from the TWG's examination of these measures are posted on the Institute for Information Law (IViR) website. In addition, our discussion also generated cross-cutting themes and insights, which we discuss below.

The members of the Transatlantic Working Group participating in the Ditchley Park Session may not necessarily agree with or endorse every observation noted below, and undoubtedly have other important ones to add to this summary. But they accept that this report reflects the main points we discussed and agreed in principle during our meeting, with the understanding that additional details and views will be reflected in subsequent publications of the Working Group.

Key findings and recommendations

We encourage policy makers, the tech industry, and other stakeholders to consider these points as they seek ways to address harmful content online without chilling free speech:

Clearly define the problems being addressed, using an evidence-based approach. Policy measures directed at vaguely defined concepts such as "extremism" or "misinformation" will capture a wide range of expression.

- Before taking any steps to restrict speech, regulators should explain clearly and specifically the harms they intend to address, and also why speech regulation is necessary for this purpose.
- The rationale should be supported by concrete evidence, not just theoretical or speculative concerns.
- Any government action should also be subjected to timely review, in order to assess whether it continues to serve its intended purpose. To this end, "sunset clauses" can be an effective tool to encourage a thorough impact review post-implementation.

Build in transparency by government and industry alike so that the public and other stakeholders can assess more accurately the impact of content moderation.

- The industry's Hash-Sharing Database, in particular, was criticized for a lack of transparency into its workings. Germany's Network Enforcement Law (NetzDG), despite other criticism, does include some transparency reporting requirements, but they need to be tightened.
- Generally, government action to direct the content moderation practices of platforms should be documented and available for academic research as well as the public.
- Platforms should also share detailed information about their content moderation practices with the public, working with the public and academics to design such disclosures or databases while respecting the privacy of the people who use their services.

Ensure due process safeguards for online speech.

- When user-generated content is removed, the authors often have limited or no redress. This practice may facilitate unwarranted censorship and abuse or perceptions of arbitrariness. The uploading user should be offered a clear and timely recourse mechanism for considering reinstatement.
- When governments direct action to restrict online speech, their measures should comply with rule of law principles so that they are subject to judicial review; governments should *not* use informal agreements with private platforms to obscure the role of the state and deprive their targets of civil redress.
- Platforms should consider notifying content providers when they receive a formal notice from government to remove that content, so that content generators can appeal the decision with the appropriate authorities. For example, the NetzDG law does not provide for appeal mechanisms, nor are users notified of official complaints levied against their content.

Reimagine the design of both public and private adjudication regimes for speech claims.

- Many online platforms already offer internal, private appeal mechanisms. However, given the democratic values at stake, the lack of judicial oversight and the resulting "privatization" of speech regulation raises concerns. Accordingly, there may be a need to create independent, external oversight from public, peer, or multistakeholder sources.
- The Transatlantic Working Group will continue to explore designs for such external review. Some options include an increased role for independent regulators, specialized judicial "online review systems," and private or multistakeholder "standards council" solutions.
- Courts should continue to play their historical role in developing a body of law through well-reasoned decisions that would provide guidance to platforms, users, and governments.

Craft appropriately tailored policies: one size need not fit all.

- Policy discussions often refer in general terms to "platforms" and/or "online intermediaries," but these concepts are too broad. They cover a wide range of services and operate at different layers of the internet stack, with entirely different abilities (and responsibilities) to moderate online speech. Policy makers should consider the different roles and capacities of these players.
  For example, content restrictions imposed at lower levels of the "stack" (such as Internet Service Providers, CDNs and the Domain Name System) have a greater impact on freedom of expression than at higher levels (such as web forums, social media, chatrooms).
- Size is another important factor: the cost of regulatory compliance disproportionately burdens smaller and nonprofit services, and should be considered when imposing requirements or penalties.
- But, a caveat: regulatory scrutiny of larger platforms has led some bad actors to migrate to smaller platforms or encrypted services, such as 8chan or Gab, where they are less likely to be removed.

Understand the risk of overreliance on automated solutions such as AI, especially for context-specific issues like hate speech or disinformation.

- Automated approaches have had some success, such as in blocking child sexual abuse content and copyrighted material. However, identifying hate speech and disinformation often requires a nuanced assessment of context and intent. While improving, automated systems still generate a significant number of false positives.
- Automated removal can act as a prior restraint, which prevents content from ever being published. Therefore, automated systems should include an adequate number of human reviewers to correct for machine error.
- AI solutions may reinforce biases, since they are trained on historical datasets that reflect broader social contexts. This can lead to unfair and biased outcomes in content moderation, and the further marginalization of certain groups. Online services should probe for and eliminate such biases. Our second and third Working Group Sessions will do a deep dive into artificial intelligence solutions.
- Given the quantity of user-generated content, automated systems necessarily are an important part of the solution. However, policy makers and industry should avoid overstating the power to solve speech problems through technical means, and should incorporate wherever possible qualitative human oversight.

Next steps

Our second Transatlantic Working Group Session in May will examine initiatives to address viral deception (disinformation), especially in the context of elections; self-regulatory models, including the European Commission's "Code of Practice"; emerging regulatory frameworks, including the British Government White Paper; practices surrounding "takedowns"; algorithms and accountability; and will introduce a discussion of intermediary liability.

In the fall, our third and final session will further examine the earlier topics and focus in depth on artificial intelligence and on intermediary liability.

Between these sessions, we will continue to reach out to diverse stakeholders and the public in roundtables and forums for their feedback and engagement.

TRANSATLANTIC WORKING GROUP

Co-Chairs Report No. 2: The Bellagio Session

Susan Ness, Annenberg Public Policy Center
Nico van Eijk, Institute for Information Law, University of Amsterdam
February 13, 2020

Introduction

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

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

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

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

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

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

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

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

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

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

Key findings and recommendations

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

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

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

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

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

The "A-B-C" analytical framework, which was presented in one of the papers and discussed during the meeting, helps policymakers to focus not on *Content* but rather on bad *Actors* and deceptive *Behavior*. As content, "fake news" and "disinformation" are part of our democratic landscape, and are not *per se* illegal. Governments should not trigger the removal of undesirable, but legal, content, as such action is inconsistent with freedom of expression.

Governments may choose to address harmful behavior on the internet, where content is artificially propagated by bad actors – "bots, "astroturfers," or "troll farms" – as such conduct no longer reflects an authentic dialogue among citizens. This distortive behavior is akin to "spam," which companies

have the technical expertise to address. But a cautionary note: some fake identities might be legitimate and should be protected, such as whistle-blowers calling attention to government corruption.

Strengthen enforcement of rules on foreign government interference

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

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

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

Strengthen transparency and accountability

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

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

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

During an election season, special attention should be given to both candidate and social issue advertising, as such communications are integral to the electoral process. If narrowly drafted, governments could require specific disclosures for microtargeted candidate and social issue ads that state why the ad is being seen, the screening criteria, who paid for the ad, and the amount spent.

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

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

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

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

Be cautious if considering changing intermediary liability laws

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

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

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

Promote media literacy, quality journalism, and fact checking

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

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

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

Good Corporate Governance Encompasses Good Corporate Citizenry

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

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

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

Next Steps

Our final Transatlantic Working Group Session in November will examine:

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

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

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

TRANSATLANTIC WORKING GROUP

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.