TRANSATLANTIC WORKING GROUP

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.