collective wellbeing Carnegie UK

A modern systems approach to regulating online hate speech

EPRA 29 June 2022

About us

Lorna Woods is Professor of Internet Law at the University of Essex and a member of the Human Rights Centre there. Lorna is well-known as a European lawyer for the Textbook on EU Law (Steiner and Woods). Lorna is a senior associate research fellow at the Information Law and Policy Centre, Institute of Advanced Legal Studies, University of London and a member of the policy network at the Centre for Science and Policy, University of Cambridge. Lorna recently co-edited 'Perspectives on Platform Regulation: Concepts and Models of Social Media Governance Across the Globe' (Recht Und Digitalisierung U Digitization and the Law, with Judit Bayer, Bernd Holznagel and Paivi Korpisaari).

William Perrin is a Trustee of Carnegie UK, one of the global family of trusts set up by Scots-American philanthropist Andrew Carnegie. William was policy advisor and Private Secretary to Prime Minister Tony Blair from 2001-2004 where he was responsible for reforming regulation of communications, de-regulation of pubs bars and clubs and regulation of gambling. William delivered the 2001 Communications White Paper which created modern communications regulator OFCOM and abolished seven predecessor regulators.

The hate speech challenge

> *"We have evidence from a variety of sources that hate speech, divisive political speech, and misinformation on Facebook and the family of apps are affecting societies around the world.* ***We also have compelling evidence that our core product mechanics, such as virality, recommendations, and optimizing for engagement, are a significant part of why these types of speech flourish on the platform . . . the net result is that Facebook, taken as a whole, will be actively (if not necessarily consciously) promoting these types of activities. The mechanics of our platform are not neutral****."*[18] (emphasis added)

Whistleblower testimony to Congress

- The systems and processes run by the platforms are fundamental to the problems
- Colossal scale means a piece by piece approach will never work

WHISTLEBLOWER AID

Report government and corporate lawbreaking. Without breaking the law.

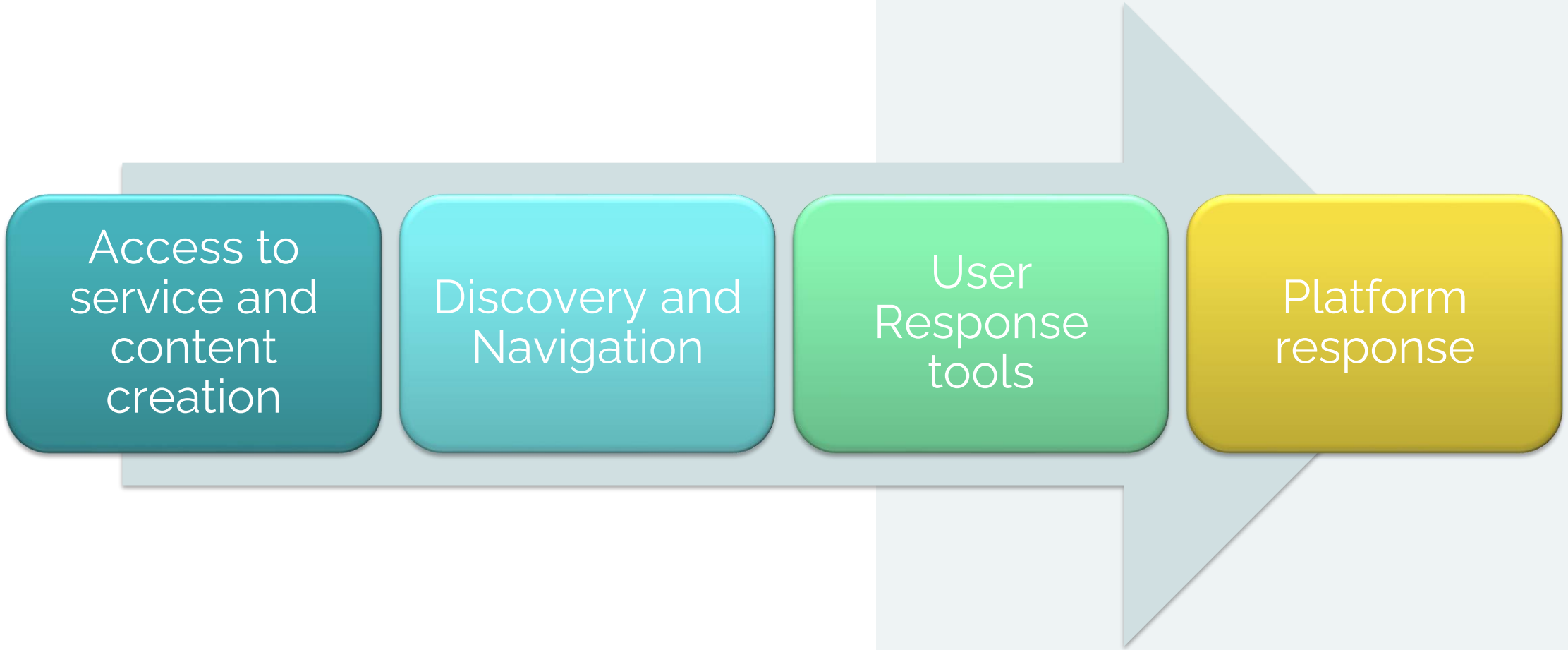
A process-driven approach to content creation and management

- Platforms are defined by their systems and processes – in software and basic rules (Lessig)
- Orientation of processes and investment in these systems and processes reflects shareholder objectives; these may have unintended consequences for content creation, dissemination and user experience
- Our approach is to use these as points for regulatory intervention

"Facebook … knows—they have admitted in public—that engagement-based ranking is dangerous without integrity and security systems but then not rolled out those integrity and security systems in most of the languages in the world,…It is pulling families apart. And in places like Ethiopia it is literally fanning ethnic violence."

Frances Haugen
Facebook Whistleblower
Testimony to US Senate

Four areas of system intervention



Design Choices - examples

- Metrics (and other hooks)
- Frictionless communication (retweet; like share; ease of forwarding)
- Content discovery systems (search engines/recommender systems)
- Clickbait rewards and ‘outrage porn’
- Targeted advertising
- But little investment in making complaints resolution easy/effective

Examples of system interventions

- WhatsApp reduction in forwarding velocity - India
- Instagram – introducing age assurance
- Twitter – 'Do you want to read article before forwarding?'
- Twitter, Facebook de-algorithming Alex Jones/Infowars
- UK OSB – special appeal process for journalists
- Google autocomplete changes 'The jews are….'

Outcomes-oriented

- Risk management model – similar to any dangerous industry
- Large systems – outcomes – when the company makes adjustments does surveyed level of harms go up or down?
- Regulator codes of practice influence how company achieves outcomes
- Harm reduction cycle

Hate speech code of practice

- Developed with representatives of victims – OFCOM and companies observed process.
- How can systems and processes be managed better to reduce hate speech
- UN Guiding Principles on Business and Human Rights
- OECD Guidelines on Responsible Business conduct
- Ruggie – both based on risk management in systems and processes
- Three pillars of Ruggie Principles: protect, respect and remedy.

Hate speech code: Guidelines in 12 areas

1. Corporate responsibility
2. Proportionate risk assessment, mitigation and remediation
3. Safety by Design
4. Access to network and content creation
5. Discovery and navigation
6. User response, user tools
7. Moderation
8. Safety testing
9. Supply chain issues
10. Victim support and remediation
11. Education and training
12. Vigilance over time

Guideline 1: Responsibility, Risk Assessment, Mitigation and Remediation

(1) Social media service providers should have a policy commitment to take action to combat hate speech arising on their service. This commitment should be endorsed by the global board and all 'c-suite' executives.

(2) Social media service providers should carry out a suitable and sufficient assessment in relation to each nation in which the social media service is used as to the risk of harm from hate speech attacks on people or groups based on their identity arising from the operation of the service or any elements of it. The risk assessment should be accompanied by a mitigation plan that addresses at least the issues raised later in these Guidelines.

(3) The risk assessment should, in particular, be carried out before the launch of any new service, any new feature, or any service or feature is made available in any new nation.

(4) Service providers should identify metrics to assess the appropriateness and success of the mitigation plan and use them to assess effectiveness of the mitigation plan regularly (at least annually) and revise the mitigation plan accordingly.

Guideline 2: Safety by Design

(1) Social media service providers should implement appropriate “safety by design” technical and organisational measures including but not limited to those detailed in these Guidelines to minimise the risks of those harms arising from hate speech and mitigate the impact of those that have arisen, taking into account the nature, scope, context and purposes of the online platform services and the risks of harm arising from the use of the service.

Guideline 4: Discovery and Navigation

(1). Social media service providers should review their recommender systems, especially their automated systems, so that they do not cause foreseeable harm through promoting hateful content, groups or other users to follow for example by rewarding controversy with greater reach, causing harm both by increasing reach and engagement with a content item.

(2) Social media service providers should consider whether the recommendation of “counter speech” is effectively supported by their systems.

(3) Social media service providers should consider the impact of autoplay functions, especially in the context of content curated or recommended by the provider. Where the service provider seeks to take control of content input away from the person through autocomplete or autoplay (see below) the provider should consider how this might affect a person’s right to receive or impart ideas.

This generic approach can be used widely

- Systems and processes are a common denominator for platform problems
- Can be applied in many different settings:
  - Against almost any legal background – tougher or looser depending on local law
  - On a wide range of topics – with straightforward alterations
- Developed code on violence against women and girls in UK
  - Already debated in parliament
- Stripped out topic specific references to create a model ‘universal’ code

What does this mean for a modern regulator?

- ✓ Strong information gathering powers to allow you to access companies internal risk assessments, risk mitigation plans and data
- ✓ Powers to make them create data for you
- ✓ OSB (UK) provides possibility of prison sentence for not providing information.
- ✓ Independent research capability to challenge data from platforms about outcomes
- ✓ Need staff who understand situation inside companies

collective wellbeing Carnegie UK

Thank you for listening!

Professor Lorna Woods OBE FRSA
William Perrin OBE FRSA

Carnegie UK
Andrew Carnegie House
Pittencrieff Street
Dunfermline
Fife, Scotland
KY12 8AW

T +44 (0)1383 721445
carnegieuk.org.uk