Presentation notes

Introduction

- Thanks and welcome
- Charles, Director, GPD
- Structure of presentation
    - Look at driving factors behind online content moderation trends
    - Examine trends and challenges with government regulation
    - Examine trends and challenges with company self-regulation

1 – Online Content Moderation

What's happening?

- Social media platforms routinely engage in the depublication, downranking, and censorship of information and/or user accounts.
- In recent years, this practice has become a matter of intense public interest.

Why?

- Social media platforms grown to scales we never could have imagined
- Many platforms now where users around the world seek, receive and impart information – from sharing pictures and messages with loved ones, to political speech, the news…
- With this has come all of the good, as well as the bad
- What is permissible online, and who decides, has become one of the most important public policy debates of our time.

The response:

- Two types of response to these challenges: government regulation and company self-regulation
- Neither of these have int. HR law and standards at the heart
- Let's examine the trends

2 – Government Regulation

Generation one of government regulation:

- Informal pressure of individuals and media
- Placing political and informal regulatory pressure on tech companies to remove contentious content (US, EU, UK)
- General provisions in cybercrime
- Erecting market barriers as a way to exert control over the presence and actions of companies (China)
- Blocking websites at the ISP level (most countries)
- Use of state requests for content takedowns based on platform terms of service and "shadow" requests through government-aligned actors (Ecuador, Ukraine)
- Developing non-transparent agreements with companies top remove online content that falls into particular categories (Blasphemy, copyright) (Turkey, Israel, Pakistan (prior to blasphemy law))

Generation two – legislative and regulatory proposals:

Three broad changes:

- Specific obligations to remove specific content
- General duties
- Removal of intermediary liability protection

Attachment of specific obligations to remove specific content

Governments around the world looking to attach obligations on companies to take down specific content. Sometimes this is clearly defined and illegal, sometimes not – Malaysia

Some examples

NetzDG

- Germany's NetzDG came into effect on January 1, 2018. It targets social media platforms such as Facebook and Twitter, and requires them to remove posts featuring hate speech or fake information within 24 hours.
- It applies existing laws that prohibit speech to the online space.
- A platform that fails to adhere to this law may face fines up to 50 million euros.
- And we are seeking fines coming in for under-reporting by Facebook of flagged content
- HR advocates were very worried about the law – due to the possibility of companies over-blocking
    - Government is now reviewing the law because too much information is being blocked that shouldn't be.
    - The Association of German Journalists has complained that social media companies are being too cautious and refusing to publish anything that could be wrongly interpreted under the law. This could lead to increasing self-censorship, possibly of information in the public interest.

French Hate Speech Law – hot off the press

- French MPs passed a law last week that is designed to fight online hate speech that will oblige social media networks to remove offending content within 24 hours and create a new button to enable users to flag abuse.
- Sites that fail to comply with the law and remove "obviously hateful" content risk fines of up to €1.25m (£1.12m). The upper house, the senate, will now examine the legislation, and could suggest amendments
- Modelled on NetzDG that came into force last year. Heavy penalties, short timelines for takedowns.
- Covers search engines and social media.
- Contains measures to tackle offences such as harassment, hate speech, pimping and condoning terrorism.

Introduction of general duties

Online Harms White Paper – DCMS and Home Office

Sets out the governments plans and will set forth the legislative framework for the UK for online content.

Explain objectives.

Challenges

Scope and Definitions of Harm

- Scope of speech includes: illegal and legal but harmful content.
- Content in scope should be clearly defined and objectively harmful.
- Although the unlawful online content which the White Paper identified as needing to be tackled is, for the most part, clearly defined in legislation, this isn't the case for the "legal but harmful" content which is also within its scope.
- So while it is relatively straightforward to identify child sexual abuse imagery, for example, the same cannot be said for speech which might amount to "coercive behaviour" or "disinformation", to take two examples from the White Paper. Indeed, such speech is generally lawful "offline", risking the creation of two different standards of speech depending on whether it is expressed in person or online.
- Furthermore, each type of harm identified in the White Paper requires a different, specific, and tailored response. What might work as an approach against terrorist propaganda would be manifestly inappropriate for cyberbullying. Yet the White Paper proposes only one blanket form of regulation for all these very different harms.
- Without definitions, and particularly if there are sanctions for non-compliance with any duties, platforms will be incentivised to interpret the terms broadly, rather than risk sanctions, and, therefore, remove an even broader range of content than is intended, including content which is protected by the right to freedom of expression.
- We are particularly concerned that this could create a perverse situation where speech which is lawful, but potentially harmful, is restricted when it is expressed online, but not when it is expressed in person

Time frame

- Timeframes are undefined in the whitepaper but suggest a challenge.
- **Short time limits risk rushed decision-making and an inability to fully consider context** or obtain the necessary information and expertise in order to make an accurate determination
- Seemingly seeks to increase proactive monitoring – privacy and FoE risks here
- Not only In violation of the e-commerce directive but creates and incentive structure to over-remove content.
- Further exacerbated by machine mediated content moderation
- The imposition of time limits would incentivise the use of automated processes for determining whether content is unlawful or harmful.
- However, automated processes are extremely poor at making determinations relating to the nature of content given their inability to determine context, and the difficulties in defining terms such as "bullying" or "insult".[2]
- Examples Tumblr, keyword search/ screening – do not appreciate context

Sanctions

Companies have to decide whether content is on or out of scope

Sanctions create strong incentive to "play it safe" and simply remove the content rather than risk a sanction

- Evidence from the implementation of the Network Enforcement Act (NetzDG) in Germany in 2018 suggests that this would be the case: since the introduction of the tight timelines and heavy fines included in the NetzDG legislation (48 hours in the case of "manifestly unlawful" content), there have been a number of high-profile examples of Twitter, for example, removing tweets which were controversial, satirical and ironic, but not obviously illegal or even harmful.[4]
- Sanctions and penalties should be imposed as a last resort

Removal of intermediary liability protections

- Section 230 of the Communications Decency Act and Articles 12-14 of the e-commerce directive
- Basis for internet intermediaries
- Interesting conflict between these second generation legislative proposals and the intermediary liability protections on the e-commerce directive. This will be reviewed by the next commission, but to what extent will be tbd.

3 – Company self-regulation

Phase one –

- Purely neutral platforms, intermediaries.
- Challenges increased and needed greater control

Phase two –

- Three types of development: substantive developments, enforcement and oversight

Substantive Developments

- Companies developing greater detail on their own ToS
    - making them clearer
    - providing examples
    - making them easier to find for users
- Industry initiatives trying to create shared substantive standards (GNI as example) but narrowly applied to government requests (not applied to ToS)
- Voluntary initiatives now applied, such as EU Code of Conduct on Disinformation

Enforcement

- Greater number of moderators (FB now 30k)
- Introduction of algorithms for content removal and downgrading (user flagging)
- Graduated and nuanced approach to the availability of content
- Now we have pre-emptive removal from platforms on particular types of content
- IWF as joint initiative for CSAI
- GIFCT as joint initiative for terrorist content – new development to watch here is Christchurch Call and Crisis Protocol (which goes beyond determining enforcement)

Oversight

- FB board the first here
- GPD has called for industry wide oversight board
- Increase accountability and transparency of standards and enforcement

4 - Conclusion

- Neither approach, government regulation or company self-regulation, is compliant with international HR standards
- There needs to be much more joined up thinking between the two
- Greater HRDD and HRIA of policy development
- Greater evidence of the harms caused and implications of policy/legal development/changes
- Greater stakeholder engagement