BUILDING TRUST

#02: HARMFUL CONTENT

A resource provided by Friend MTS

BUILDING TRUST

Platform responsibility and online harms have become big talking points among policy makers and the political community in recent years.

While the debates continue, it's clear that the regulation of online platforms is going to increase in the years ahead, and that the corporate reputation of digital companies is going to increasingly depend on well-thought-out and communicated policies in this domain.

It's important that platforms, rights-holders and responsibility facilitators are all well informed to ensure and enable productive discussions around these topics. But understanding many of these issues means navigating often complex areas of law and contentious social debates.

To help inform and educate all stakeholders, Friend MTS has teamed up with media and music consultancy CMU Insights to present Building Trust, a series of white papers exploring the economic and social responsibilities of digital platforms and other online service providers.

In this second white paper we look at the debate around harmful content – which is to say offensive, unlawful, abusive and misleading content.

We consider what digital platforms currently have to do and choose to do when content of this kind is posted by their users. And we review various proposed legal reforms around the world that will increase the responsiblities of platforms when it comes to dealing with the different kinds of harmful content that is routinely uploaded by users to their networks and servers.

Content from the Building Trust Series may be shared but must reference Friend MTS as the originator and cite the article from where it is pulled. For more information on redistributing this content, please reach out to pr@friendmts.com.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

HARMFUL CONTENT

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

ABOUT THE AUTHOR

This white paper has been compiled by **Chris Cooke**, Founder and Managing Director of London-based media and music consultancy CMU Insights. As a business journalist, consultant and educator, Chris has been writing and talking about the media, music, digital and copyright industries for 20 years.

He monitors current trends and best practice, and helps people and companies working in these sectors to navigate and understand the latest developments, challenges and opportunities.

ABOUT THE INTERVIEWEES

This white paper is based on a series of interviews and a review of recent articles undertaken between May and August 2021 via which we consulted the following...

Reed Smith is a leading global law firm. We quote **Carolyn E Pepper**, a partner at Reed Smith whose specialisms include media, social media, media regulation and intellectual property.

First Draft is an organisation with bases in London, New York and Sydney that aims to "protect communities from harmful misinformation". We quote its Head Of Impact & Policy **Tommy Shane**.

The **Open Rights Group** is a UK-based digital campaigning organisation that aims to "protect our rights to privacy and free speech online". We quote Policy Manager **Heather Burns** and Executive Director **Jim Killock**.

The **Electronic Frontier Foundation** is a US-based campaigning organisation that aims to "defend civil liberties in the digital world". We quote the EFF team and one of the group's special advisors, **Cory Doctorow**.

The **World Federation Of Advertisers** is a global organisation for the advertising and marketing sectors, and founder of the Global Alliance For Responsible Media. We quote its CEO **Stephan Loerke**.

We also quote from the UK government's **Online Harms White Paper** (published December 2020) as well as the **UK Online Safety Bill** (published May 2021) and **EU Digital Services Act** (published December 2020).

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

SECTION 1

INTRODUCING HARMFUL CONTENT

The obligations of internet companies and digital platforms in relation to so called "harmful content" or "online harms" have become a big talking point within the wider media and political community in recent years – with headline-grabbing incidents on social media and other platforms regularly putting the spotlight back on this debate.

The internet – and the rise of social media and user-upload platforms – have provided powerful communication and publishing tools for the wider population. But what happens when people use those tools to cause "harm"? And to what extent should it be the responsibility of the platforms to regulate and restrict any harmful content as it passes through their networks?

Law-makers across the world have started to discuss these questions – both formally and informally. In some countries, politicians are currently actively considering proposals that internet companies and digital platforms should have increased obligations to reduce online harms, mainly by monitoring, filtering and/or removing harmful content.

The same companies and platforms have also come under increased pressure from their users and advertisers in recent years to take a more proactive, consistent and transparent approach to dealing with both harmful content and the people who create and post it.

However, dealing with content of this kind is both challenging and controversial. For starters, it can be very hard to even define what is meant by "harmful content". And, in many cases, whether or not content is "harmful" will depend on the context in which it was created and published, and possibly also on the opinions and perceptions of the audience.

A distinction can possibly be made between that content which actually breaks an existing law – either through its creation or its intent – and that content which might be considered harmful but which isn't technically illegal. Should the priority be restricting the former, ie overtly illegal content?

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

But then, is it for digital platforms to ascertain whether a piece of content breaks any laws, or is that only a role that can be performed by law enforcement and/or the courts? Plus some would argue that platforms should also be restricting and removing content that isn't technically illegal, but which nevertheless causes harm.

Removing harmful content also raises important concerns around freedom of expression. This will depend to an extent on the level of harm being caused, but where the outcome is simply that an audience is offended or misled, is that of sufficient concern to remove such content and, in effect, to infringe on the free speech rights of the creator?

Meanwhile, simply identifying harmful content in the first place may cause additional concerns around data protection and privacy. If platforms are to face any new obligations regarding online harms, said obligations need to factor in any existing regulations designed to protect a user's data.

And even if they do, some platforms will argue that their technologies actually go beyond current legal obligations when it comes to data security, and those enhanced efforts could be negatively impacted if there are new requirements to monitor all content for potential harms.

Despite these challenges, many countries are now proposing new laws that would oblige at least the biggest digital platforms to more rigorously regulate and restrict harmful content online.

In the UK, an Online Safety Bill was presented to Parliament in May 2021. In the European Union similar proposals are being considered as part of the Digital Services Act initiative. In the US, there have been various debates in Congress, and beyond, about reforming Section 230, the current law that restricts the liabilities of digital platforms in this domain.

Meanwhile, the big digital platforms have all been evolving their own policies and systems for dealing with harmful content. Partly to pre-empt and possibly lessen the need for new obligations under law. And partly to protect their corporate reputations, which are increasingly under pressure in this domain, and to placate increasingly concerned users and advertisers.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

In this white paper we will seek to define harmful content and to organise the different kinds of content that arguably cause some sort of harm. We will then look at how both law-makers and digital platforms around the world are currently dealing with these issues, and the concerns that have been raised about those efforts, particularly in relation to free speech.

SECTION 2
DEFINITIONS

The terms "harmful content" and "online harms" are sufficiently ambiguous that they could cover a wide range of online activity and content.

Some have argued that economic harm should also be part of this conversation, which – among other things – would bring intellectual property concerns into the debate. After all, when people use social media and user-upload platforms to distribute copyright infringing material, the original content creator and copyright owner is being economically harmed.

However, in the main copyright matters are not part of the online harms debate in political circles. Reforms to copyright law – and especially the copyright safe harbour that restricts the copyright liabilities of digital platforms – are being actively considered in multiple countries of course, but separately to this debate.

Concerns around data and privacy could also be considered under the banner of online harms – which is to say how digital platforms collect, protect and exploit user data.

Where user data is not utilised in a responsible and transparent way – and especially if data is not stored securely and becomes available to rogue entities – then clearly the users whose data has been gathered can be economically harmed. But again, this is not really part of the online harms debate, with the obligations and liabilities of digital platforms when it comes to protecting user data and privacy covered by separate legislation.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

That's not to say that certain lobbying groups – especially within the copyright industries – haven't tried to have economic harms of this kind included in the online harms debate. And some of the proposed legal reforms being made on the back of that debate – especially around transparency – could be beneficial to copyright owners too.

However, in the main the formal proposals being made in relation to 'harmful content' do not overtly cover intellectual property or data concerns.

So, what are we dealing with? Well, we can organise the kinds of content that are usually included under the online harms banner into at least four main groups – those would be: offensive content, unlawful content, abusive content and misleading content.

OFFENSIVE CONTENT

This is simply content that offends certain people. It includes content of a sexual or violent nature; content that includes words that some people consider offensive; controversial political opinions; and content that mocks or attacks specific people or groups of people, especially when that is based on things like race, nationality, religion, sexuality or gender.

Within the conventional media – especially broadcast media – this kind of content has traditionally been regulated to an extent. Some – especially content that which is racist, misogynistic or homophobic, or more extreme sexual or violent content – might be banned outright.

With other offensive content, the rules may mainly aim to ensure that it is not accessed by children, and that an adult audience is given advance warning that the content they are about to consume may be considered offensive in one way or another.

Even where broadcast media rules do actually ban offensive content outright, that content is still often legally available via other channels for those who wish to access it.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

HARMFUL
CONTENT TYPES

OFFENSIVE CONTENT
CONTENT THAT OFFENDS

adult content

pornographic content

violent content

offensive language

offensive opinions

UNLAWFUL CONTENT
CONTENT THAT IS ILLEGAL TO CREATE

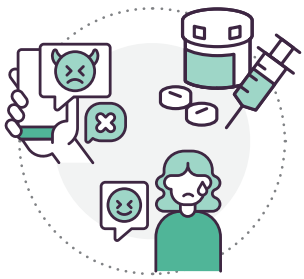
child sexual exploitation and abuse

terrorist content and activity

incitement of violence

promotion of illegal goods

extreme pornography



ABUSIVE CONTENT
CONTENT THAT SETS OUT TO ABUSE

cyber-bullying

trolling

revenge pornography

encouraging or assisting suicide

advocacy of self-harm

MISLEADING CONTENT
CONTENT THAT MISLEADS

disinformation

misinformation

propaganda

conspiracy theories

fake news

Of course, in some countries, there is still actual censorship of at least some offensive content, either as a result of specific laws or because of voluntary codes signed up to by the media and entertainment industries. However, in many countries censorship of this kind has been phased out in favour of age restrictions and viewer advisory warnings.

Free speech rights are frequently cited in opposition to the censorship of offensive content. The right to freedom of expression under law is usually subject to some restrictions, but many would argue that people being offended is not a sufficient justification to restrict free speech. The extent to which the law agrees does vary from country to country, and according to the specific way, and degree to which, any one piece of content offends.

UNLAWFUL CONTENT

This is content that it is actually unlawful to create in the first place. While such content is not new, the rise of digital channels and platforms have made it much easier to access and distribute content of this kind, in turn resulting in more such content being created.

Two key categories of unlawful content are frequently discussed as part of the wider online harms debate within political circles – those being all and any forms of content linked to child sexual exploitation and abuse (CSEA) and terrorist activity.

On the former, a 2019 white paper from the UK government on online harms described how "child sex offenders use the internet to view and share child sexual abuse material (CSAM), groom children online, and live stream the sexual abuse of children".

On the latter, the same paper said that "terrorists – including Islamist groups such as Daesh and Al-Qaeda as well as far right terrorists – use the internet to spread propaganda designed to radicalise vulnerable people, and distribute material designed to aid and abet terrorist attacks. There are also examples of terrorists broadcasting attacks live on social media".

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

Other kinds of content may also be considered unlawful simply because it encourages or facilitates other crimes, which may occur online or offline, such as inciting violence.

The creation of extremely violent content – and more extreme forms of pornography – may also be illegal, usually as a result of laws that seek to protect those who might be involved in the creation of such material, even where those people, in theory, consent.

Where the creation of content is in itself illegal, the distribution of such content is also likely to be unlawful, as may be the possession and consumption of these materials.

ABUSIVE CONTENT

This is content that sets out to attack, harass, bully or intimidate specific individuals, or small groups of individuals, commonly referred to as cyber-bulling, cyber-stalking or trolling.

This content is usually highly critical of, or threatening towards, specific individuals. Where it is simply highly critical there is a blurring of the lines between offensive and abusive content.

Which is to say the content in itself is simply offensive, though the way in which it is delivered could be seen as abusive. Obviously if threats are being made it is clearly abusive, although whether or not something is actually threatening may be open to interpretation.

The person creating this content may simply post it to their own profile, page or channel on a social media or user-upload site. Or they may seek to ensure that the person they are targeting sees the content. There are various ways that can be achieved: eg posting a comment on the targeted person's profile; sending a direct message; or simply including the person's user-name or handle on a platform where doing so alerts the subject of the post.

There are other types of content that could also sit under this 'abusive content' category.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

The first type is content that was never intended for public consumption, but which is nevertheless made public, often with the intent of embarrassing, humiliating or intimidating another person. This would include the publication of confidential documents, information or correspondence – often referred to as 'doxing' – and also the posting or distribution of intimate sexual images online by an ex-partner – often referred to as 'revenge pornography'.

The second type is content that seeks to influence other people in a way that encourages them to harm themselves or to commit a crime. This content could be generally accessible online or be targeted at specific individuals through social media or messaging apps. One example that is often referenced is content that advocates or explains methods of self-harm or suicide.

The posting and distribution of abusive content may be unlawful, with many of the scenarios described here variously covered by either specific laws relating to internet usage and communication, or general laws dealing with harassment, blackmail or extortion.

However, unlike with the unlawful content category described above, the actual creation of the abusive content in itself is often not unlawful. Rather it is the way in which the content is published, distributed or targeted at specific individuals that makes it illegal.

MISLEADING CONTENT

The final core category of potentially harmful content is disinformation and propaganda designed to persuade or mislead individuals, and/or to manipulate public opinion.

This kind of content is often colloquially described as "fake news", although that is a term that has been variously used to cover a wide range of things, from biased journalism to political spin to clickbait articles and videos, to outright lies and meritless conspiracy theories.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

HARMFUL
CONTENT TERMS

One of the challenges when discussing 'harmful content' is that there isn't really any consensus on how we should refer to the different kinds of content that can cause harm. We reviewed the policies and guidelines of the big platforms to identify the kind of language they employ when talking about the four categories of harmful content.

OFFENSIVE CONTENT

- "objectionable content"
- "harmful content"
- "inappropriate content"
- "offensive content" "sensitive media"
- "sensitive content"

UNLAWFUL CONTENT

- "violent content"
- photos depicting "anything overtly pornographic or children's naked bodies"
- "violent or graphic content"
- "graphic violence" and "violent sexual conduct"
- "violent, graphic or shocking content"

HARMFUL
CONTENT TERMS

ABUSIVE CONTENT

“hate speech” and “abuse”	
“harassment”	
“harrassment” and “profane content”	
“hateful content” and “online abuse”	
“harmful or dangerous content”	

MISLEADING CONTENT

“false news”, “inauthentic content” and “clickbait”	
“false information”, “fake accounts” and “fake and misleading user reviews or ratings”	
“misinformation”, “fake accounts” and “spam”	
“inaccurate”, “inappropriate”, “mislabeled”, “deceptive”, “spam” and “fake accounts”	
“unwanted or mass solicitations (spam)”, “deceptive practises”, “fake engagement” and “impersonation”	

The trademarks for the platforms referred to above are the property of their owners.
Friend MTS is not connected with, and does not endorse, any of those platforms.

Disinformation and propaganda are not new, of course, although many would argue that digital channels have allowed such content to become more widespread, while algorithms on social media and search engines make it easier for the creators of this content to target their messaging according to an audience's pre-existing biases and prejudices.

Specific concerns about misleading content have also been raised during the COVID-19 pandemic, with fears that the spreading of disinformation can also affect public health.

This is another area where free speech is central to the debate. Misleading content may be distributed and targeted in a way that interferes in a country's democratic processes. But freedom of speech is a core principle of democracy too. How do you go about regulating harmful misleading content without infringing on people's free speech rights?

SECTION 3
LEGAL PRESSURE

In most countries, the responsibilities of internet platforms, when it comes to dealing with harmful content, are currently restricted, with the voluntary codes of the big social media and user-upload platforms often going significantly beyond their actual legal obligations.

> **"At the moment, social media is regulated by the same laws as apply to other entities, for instance the laws relating to libel, pornography, defamation, terrorism and so on. However, there are also exceptions to liability for social media platforms if they take down material once they are informed it is there. Traditional media, unlike social media to date, is also subject to other regulation – in the UK by Ofcom – and so is more heavily regulated than social media"**
>
> *Carolyn E Pepper, Partner, Reed Smith*

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

Obviously, where content is in itself illegal – or abusive content is posted or distributed in such a way that it is unlawful – the people involved in the creation and posting of that content may be investigated and prosecuted by law enforcement.

Digital platforms and internet companies may be required to assist in those criminal investigations, especially where people are posting or distributing content anonymously.

However, in most countries platforms have limited liabilities when it comes to proactively removing or blocking harmful content, even when that content is unlawful. Internet platforms are being treated more like postal services or telephone companies than broadcasters, in that they are not legally responsible for the content that flows through their networks. But there are plenty of people who think that they should be.

Digital platforms may be obliged to investigate and potentially remove allegedly harmful content – in particular unlawful content, but possibly also some kinds of abusive content – whenever they are formally made aware of it, either by users or a government agency.

Quite how this works, and how far the platforms have to go to facilitate harmful content removal, differs from country to country.

However law-makers in multiple jurisdictions are now looking into increasing the obligations of digital platforms in this domain. That might include expanding the kinds of content where such obligations apply – especially where children have access – or introducing more specific or extensive requirements in how harmful content removal systems work, which will need to be met to avoid liability.

All of those obligations could be extended even further, so that platforms must actively monitor and block at least some forms of harmful content as and when it is uploaded.

Whenever any new obligations of this kind are proposed, free speech and privacy concerns will be raised. As a result, it could be that law-makers also introduce another set of obligations forcing platforms to consider both of these things when assessing or monitoring allegedly harmful content, which might then also apply to voluntary measures.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

ONLINE SAFETY BILL IN THE UK

The UK government published a white paper discussing possible new regulations in relation to online harms in September 2019. This, in turn, informed proposed new legislation, the Online Safety Bill, which was presented to Parliament in May 2021.

Setting out the objectives of UK ministers, the 2019 white paper ran through an assortment of online harms before stating: "There is currently a range of regulatory and voluntary initiatives aimed at addressing these problems, but these have not gone far or fast enough, or been consistent enough between different companies, to keep UK users safe online".

It then added: "Many of our international partners are also developing new regulatory approaches to tackle online harms, but none has yet established a regulatory framework that tackles this range of online harms. The UK will be the first to do this, leading international efforts by setting a coherent, proportionate and effective approach that reflects our commitment to a free, open and secure internet".

As originally proposed in the 2019 white paper, the Online Safety Bill will introduce a new 'duty of care' for both search engines and platforms that enable the user-to-user exchange of content or communication. This includes social media and user-upload platforms. It's estimated that these new obligations could apply to as many as 24,000 internet companies.

Under that duty of care, affected platforms will need to ensure that "what is unacceptable offline will also be unacceptable online".

They will also be obliged to "consider the risks their sites may pose to the youngest and most vulnerable people" and to "protect children from inappropriate content and harmful activity".

And also to "take robust action to tackle illegal abuse, including swift and effective action against hate crimes, harassment and threats directed at individuals and keep their promises to users about their standards".

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

"The draft Online Safety Bill seeks to impose several duties of care rather than one – they are duties to undertake risk assessments and duties with regard to content on services that is illegal, harmful to children and harmful to adults. The bill imposes additional requirements on services which are likely to be accessed by children and 'category one' services. Although it is not clear which services will fall within the latter, category one services are likely to be the largest and most popular social media platforms. Category one services will have obligations to carry out adult risk assessments and will have additional obligations regarding privacy, freedom of expression, journalistic content and content which is harmful to adults"

Carolyn E Pepper, Partner, Reed Smith

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

The proposed legislation organises platforms affected by the new rules into two groups, and also distinguishes between the impact of harmful content on children and adults. The bill also deals with user-to-user services and search engines separately.

What are identified as 'category one' services – currently defined as "the largest and most popular social media sites" and likely to include Facebook, Instagram, Twitter, TikTok and YouTube – will have more obligations than other services.

In particular, while all platforms will have a duty to protect children from harm, category one services will have additional obligations to also protect adults. They will also have more obligations when it comes to abusive content and misleading content, in addition to unlawful content.

The bill splits this new duty of care into several more defined duties, which include obligations to undertake risk assessments, to clearly explain how users can report harmful content on any one platform, and to keep written records of all this activity.

Arguably the most important new requirements are the following...

All affected platforms will be required to employ "proportionate systems and processes" to minimise the "presence and dissemination" of illegal content. Additionally, "where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way", it must "swiftly take down such content".

Meanwhile certain specific platforms may also be required to "use accredited technology" to identify CSEA and public terrorism content and to "swiftly take down that content – either by means of the technology alone or by means of the technology together with the use of human moderators to review [the] content identified by the technology".

The bill does acknowledge that there is a balancing act to tackling harmful content while not infringing on the free speech or privacy rights of users. To that end there are additional duties around these areas, though mainly to the effect that platforms must simply consider the impact on free speech and privacy when fulfilling the other obligations.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

"As it has been drafted, the Online Safety Bill gives sweeping powers to the Secretary Of State For Digital, Culture, Media and Sport, and potentially to the Home Secretary, to make unilateral decisions, at any time they please, as to what forms of subjectively harmful content must be brought into the scope of the bill's content moderation requirements. Shockingly, it allows them to make those decisions for political reasons. In other words, a government minister will have the authority to direct an – allegedly – independent regulator to modify the rules of content moderation on topics which are entirely subjective, entirely legal, and entirely political, and to order that regulator to enforce those new rules ... You don't have to be a policy expert, or a lawyer, to see how these illiberal powers could be misused and abused. The clauses allowing government to politicise the boundaries of legal free speech have no place in this bill, or indeed, in any bill"

Heather Burns, UK Open Rights Group

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

Although, again, there are increased obligations in this domain for the category one services, including extra requirements to protect "journalistic content" and "content of democratic importance", and also to safeguard "a diversity of political opinion".

It has to be said that the current draft of the bill has been widely criticised by all sides. Those who support increased regulations around harmful content say that the measures do not go far enough or are too vague to make any tangible difference.

Representatives for the tech sector and internet rights groups say the new obligations could prove to be too onerous, do not do enough to protect free speech, and could make it impossible for digital platforms to guarantee the privacy of their users, basically making end-to-end encryption illegal.

Those critics have also expressed concern about the powers that the proposed legislation grants to relevant government ministers and regulators, and the fact that the way those powers are exercised will ultimately decide quite how wide-ranging these new regulations prove to be.

For example, a lot will be left to the UK's media and communications regulator OfCom, which will be responsible for publishing guidance on how the new duties should work in practice. Some have expressed concerns that far too much of the detail is being left to the regulator.

Some critics, like the Open Rights Group, also argue that, not only might the new regulations be too onerous and impact on free speech, but they are not likely to be effective solutions to the problems politicians have identified.

For example, when England football players faced racist attacks on digital platforms during the Euro 2020 tournament, some argued that this kind of harmful online activity demonstrated the need for regulation and should be addressed by the Online Safety Bill.

However, the Open Rights Group countered that other laws already exist to deal with those posting racist attacks, and that those laws are better tools to deal with harm of that kind than social media regulation.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

"The wave of racist comments on social media that has followed the England football defeat is reprehensible: but so is the fiction that this problem must be 'solved' by social media companies. Government alone can ensure that the law is enforced, and see that racists are prosecuted. Making this entirely Facebook and Twitter's problem to solve is simply an abdication of responsibility ... The Online Safety Bill's approach continues to perpetuate the myth that we cannot deal with criminals online. In the case of racist football 'fans', this is extraordinarily incorrect. Most will not be trying to hide who they are, and will be easily identifiable through routine data requests. A visit from a police officer, some prosecutions and in other cases cautions, would go a long way to dealing with this problem in a way that social media regulation never can"

Heather Burns & Jim Killock, UK Open Rights Group

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

As a result of all these concerns, all sides have been lobbying hard since the Online Safety Bill was published. It remains to be seen if the reach of the new laws are expanded or reduced in the final draft that is passed by Parliament, or if any of the ambiguities are addressed.

DIGITAL SERVICES ACT IN THE EUROPEAN UNION

The Digital Services Act, and the accompanying Digital Markets Act, are a set of proposals that will increase the responsibilities of digital platforms across the European Union. This will be achieved through the adoption of new EU regulations which, if passed by the European Parliament and EU Council, would have direct effect across the union.

Initial proposals were published by the European Commission in late 2020 after a public consultation. Those proposals are now working their way through the legislative process, being discussed by multiple committees in the Parliament and by national representatives in the Council.

It's the Digital Services Act which, among other things, sets out to increase the responsibilities of digital platforms in relation to harmful content. Although, in many ways, the initial proposals do not go as far as the UK's Online Safety Bill, both in terms of the kinds of harmful content covered and the new responsibilities platforms will face.

> **"In some respects the EU Digital Services Act goes further than the UK Online Safety Bill, for example, the requirement to use independent auditors to determine compliance and additional obligations relating to transparency regarding advertising and algorithms. The EU is adopting an approach with more defined obligations than the proposed Online Safety Bill"**
>
> *Carolyn E Pepper, Partner, Reed Smith*

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

> "One of the core risks of legislating social media companies is that only the best-resourced companies will be able to afford to implement them, whether that's from employing huge armies of moderators to review content, or teams of lawyers to defend themselves against countless lawsuits. This would likely entrench the major platforms like Facebook, and make competitors, or simply smaller-scale alternatives, even more difficult to emerge"
>
> *Tommy Shane, Head of Impact And Policy, First Draft*

The new responsibilities in the DSA generally relate to what is termed "illegal content". That is defined as "any information, which, in itself or by its reference to an activity, including the sale of products or provision of services, is not in compliance with Union law or the law of a member state, irrespective of the precise subject matter or nature of that law".

However, content that is harmful, but not necessarily illegal, is not covered by the DSA. In an accompanying summary, the EC stated that "there is a general agreement among stakeholders that 'harmful' (yet not, or at least not necessarily, illegal) content should not be defined in the Digital Services Act and should not be subject to removal obligations, as this is a delicate area with severe implications for the protection of freedom of expression".

In terms of the new responsibilities, the main focus is an obligation to remove illegal content when made aware of it by government agencies or users, with "hosting services" obliged to "put mechanisms in place to allow any individual or entity to notify them of the presence on their service of specific items of information that the individual or entity considers to be illegal content. Those mechanisms shall be easy to access, user-friendly, and allow for the submission of notices exclusively by electronic means".

However, there will not be a monitoring obligation, even for unlawful content like CSEA and public terrorism materials.

Indeed, the EC's summary notes that "the new regulation prohibits general monitoring obligations, as they could disproportionately limit users' freedom

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

of expression and freedom to receive information, and could burden service providers excessively and thus unduly interfere with their freedom to conduct a business".

The Digital Services Act also provides more detail in terms of the measures that platforms must take in order to ensure that freedom of expression is respected when platforms are investigating and removing allegedly "illegal content" that they have been alerted to.

The draft regulation says that, where a platform decides to remove or disable allegedly illegal content, it must inform the person who uploaded said content, "at the latest at the time of the removal or disabling of access, of the decision and provide a clear and specific statement of reasons for that decision".

It will also provide that person "for a period of at least six months" after the content is removed "access to an effective internal complaint-handling system, which enables the complaints to be lodged electronically and free of charge".

Like the Online Safety Bill, the DSA also organises platforms into different groups, with certain groups having more obligations than the others. The group with the most new obligations are referred to as the 'very large online platforms', which are defined as services that have in excess of 45 million users within the EU.

One other interesting feature of the DSA is the inclusion of "trusted flaggers" who must be given priority by digital platforms whenever they submit notices about illegal content. Trusted flaggers will be identified in each EU country by whichever authority has responsibility for the implementation of the DSA, known as the 'digital services coordinator'.

A trusted flagger, the draft regulation says, must have "particular expertise and competence for the purposes of detecting, identifying and notifying illegal content"; represent "collective interests"; be "independent from any online platform"; and must carry out its activities "for the purposes of submitting notices in a timely, diligent and objective manner".

It is worth noting that, while the DSA does not go as far as the Online Safety Bill in terms of the kinds of harmful content covered and the extent to the platforms need to proactively deal with content, there are some areas in which the EU proposals include more obligations than the UK proposals.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

That includes things like a requirement to use independent auditors to determine compliance and additional obligations relating to transparency regarding advertising and algorithms.

The DSA continues to be debated and negotiated within the European Union, with several committees in the European Parliament discussing the proposals and making recommendations. Among other things, MEPs have been discussing how harmful content should be defined, the specific requirements in terms of content removal, and various measures to protect the privacy and free speech rights of people posting to digital platforms.

SECTION 230 DEBATE IN THE US

In the US, the liability of digital platforms regarding the content they host and distribute is heavily restricted by Section 230 of the Communications Act, added into American law in 1996 as part of a legislative package known as the Communications Decency Act.

The crucial line of Section 230 is that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider". There are some limitations on the protection provided by Section 230, however, with platforms still obliged to remove material that is illegal on a federal level.

Among the specific limitations, platforms have obligations under the Digital Millennium Copyright Act if they wish to avoid liability for copyright infringement, as described in the first Building Trust white paper. And in 2018, Congress passed the Stop Enabling Sex Traffickers Act (SESTA), which introduced a new rule that said platforms are required to remove material violating federal and state sex trafficking laws.

However, in recent years there has been much debate within political circles in the US as to whether Section 230 should be further restricted or reformed. Some have proposed excluding other kinds of harmful content from Section 230 protection, so that platforms would be obliged to remove that material.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

"Though there are important exceptions for certain criminal and intellectual property-based claims, Section 230 creates a broad protection that has allowed innovation and free speech online to flourish. This legal and policy framework has allowed for YouTube and Vimeo users to upload their own videos, Amazon and Yelp to offer countless user reviews, Craigslist to host classified ads, and Facebook and Twitter to offer social networking to hundreds of millions of internet users. Given the sheer size of user-generated websites ... it would be infeasible for online intermediaries to prevent objectionable content from cropping up on their site. Rather than face potential liability for their users' actions, most would likely not host any user content at all or would need to protect themselves by being actively engaged in censoring what we say, what we see, and what we do online. In short, [Section] 230 is perhaps the most influential law to protect the kind of innovation that has allowed the internet to thrive since 1996"

Electronic Frontier Foundation

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

Others have said there should be new obligations for platforms to meet in order to enjoy Section 230 protection in the first place. And others still have proposed that Section 230 should be completely replaced with a new regulatory regime.

Both Republicans and Democrats in Congress – and both Presidents Donald Trump and Joe Biden – have proposed Section 230 reforms at various points, although the motivations, priorities and proposals of each group often differ.

Some have proposed new obligations on platforms to remove unlawful content, abusive content and/or misleading content, or at least new obligations for platforms to clearly explain their policies regarding this content and its removal. But some proposals move in the other direction.

Free speech concerns are frequently raised, as you'd expect. And in some cases proposed Section 230 reforms have begun with free speech issues, rather than proposed new obligations for platforms around harmful content causing concerns regarding freedom of expression.

Some have criticised the voluntary systems already employed by some of the big platforms for removing harmful content of one kind or another, arguing that those systems violate the free speech rights of users under the First Amendment. Such claims are usually accompanied by complaints that digital platforms display political bias when dealing with and removing certain kinds of allegedly harmful content.

Section 230 also currently provides platforms with some protection in this domain as well. It says "no provider or user of an interactive computer service shall be held liable on account of any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected".

Those Section 230 critics that begin with free speech concerns often argue that that particular protection should be narrowed somewhat.

Various proposals for Section 230 reform have circulated in Congress. The proposal that has progressed the most is probably the Eliminating Abusive And Rampant Neglect Of Interactive Technologies Act – or the EARN IT Act – which is focused on combating CSEA content.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

"Unfortunately for all of us, many of the people who don't like big tech's moderation think the way to fix it is to eliminate Section 230, a law that promotes users' free speech. Section 230 is a rule that says you sue the person who caused the harm while organisations that host expressive speech are free to remove offensive, harassing or otherwise objectionable content. For the first time, there's a law before Congress that could make big tech more accountable and give internet users more control over speech and moderation policies. The promise of the ACCESS Act is an internet where if you don't like a big platform's moderation policies, if you think they're too tolerant of abusers or too quick to kick someone off for getting too passionate during a debate, you can leave, and still stay connected to the people who matter to you. Killing Section 230 won't fix big tech. The ACCESS Act won't either, by itself – but by making big tech open up to new services that are accountable to their users, the ACCESS Act takes several steps in the right direction"

Cory Doctorow, Electronic Frontier Foundation

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

Meanwhile, in 2020, under the Trump presidency, the Department Of Justice also made a number of proposals for reform, including extending the limitations created by SESTA to other kinds of unlawful and abusive content, including CSEA and terrorism content, and cyber-stalking.

There were also proposals on the free speech front too, that being a particular bugbear of Trump. In particular, the list of material that platforms could block without liability would be amended to remove the somewhat broad "otherwise objectionable" category.

More recently, proposed reforms back in Congress have focused on public health, and the kind of misinformation that has circulated online during the COVID-19 pandemic, and also the role of social media algorithms in pushing content to users that cause physical or emotional harm.

All that said, while Section 230 reform is by no means only a Republican concern – with current President Biden also overtly criticising the law during the last Presidential election campaign, and Democrats behind some of the proposals in Congress – getting the law actually changed, and increasing the responsibilities of digital platforms, does not seem to be quite so big or urgent a priority for the current Biden administration.

There are plenty of defenders of Section 230 in its current form.

The Electronic Frontier Foundation, for example, argues that the protections it provides to digital platforms are incredibly important, and that any major changes to the principle would have a signficiant impact on the internet and how we use it. They also argue that there are often better ways to deal with many of the concerns raised by those who promote Section 230 reform, such as making it easier for users who object to any one platform's harmful content policies to leave that platform.

Of course, technically it is easy for users to walk away from the digital platforms they use, except that the dominance of certain platforms and the extensive data they gather in relation to each user makes switching services a bigger deal than it should be. Therefore the EFF points to proposals such as the ACCESS Act – which seeks to force the biggest social media companies to make it easier for users to move from one platform to another.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

"There will always be more that platforms can and should do, and one of the greatest risks is complacency. Behaviours and technologies evolve, and bad actors innovate, so this will always be a work in progress. What we now need to see more of is platforms investing in human content moderation with the necessary support and respect for the people doing that difficult work. We also need to fundamentally rethink what behaviours platforms should be optimising for. Instead of reactive, high-speed, emotion-fuelled interactions, we need more thoughtful, slower interaction with more indicators about credibility and provenance to empower all of us when we're online. Perhaps most fundamentally, we need to achieve greater consensus on what speech regulation looks like in the modern world, and how this should apply to platforms' content moderation policies. We still have a long way to go"

Tommy Shane, Head of Impact And Policy, First Draft

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

SECTION 4

PLATFORM POLICIES

All big social media and user-upload platforms have lengthy guidelines and policies regarding how they deal with harmful content on their networks. These are informed partly by existing legal obligations, although often go further than what the law currently demands due to the commercial necessity to provide an environment where the majority of users feel safe to post and consume content, and where advertisers are happy to spend money.

Most of the big platforms have some sort of automated filtering for at least some kinds of harmful content, while also relying to an extent on users reporting content that they deem to be offensive, unlawful, abusive or misleading. Content flagged that way is then usually checked by a team of moderators who decide whether content should be blocked or removed, and whether action needs to be taken against whoever uploaded the content.

Policies may differ between offensive, unlawful, abusive or misleading content. For example, unlawful and abusive content may be removed outright. But simple advisory warnings might be placed in front of offensive content, while misleading content may be flagged with links to theoretically unbiased information about the topic being discussed.

> **"When it comes to new laws requiring 'proportionate systems and processes' to minimise the 'presence and dissemination' of illegal content, many of the larger platforms are already voluntarily following codes of practice, adopting their own procedures in order to deal with these issues and adopting technology which will assist them to do so – so this requirement is not of itself likely to cause an issue for the big platforms"**
>
> *Carolyn E Pepper, Partner, Reed Smith*

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

PLATFORM COMPARISON

What do the big platforms currently say about harmful content? We reviewed the policies and guidelines of five big platforms to see how they talk about harmful content and how they deal with harmful content uploaded by their users. We also assessed how easy these platforms make it for users to access and understand this kind of information through both their web-based and app interfaces, giving each platform a score out of five.

					
Approximate # of Active Users	2.85 Billion	1.38 Billion	756 Million	397 Million	2 Billion
Key Policies	community standards / terms and policies	community page / community guidelines / terms of use	professional community policies / user agreement	terms of service / user agreement/ rules and policies	terms / rules and policy / community guidelines
Accessibility to user guidelines	●●○○○	●●●○○	●●●●●	●●●●●	●●●●●
Process & Decision Makers	user reporting / AI content filters / moderation team / oversight board	user reporting / AI content filters / moderation team	user reporting / AI content filters / moderation team	user reporting / AI content filters / review board	user reporting / AI content filters and copyright moderation / review team
Actions	1. The user receives a warning. 2. Account restricted or disabled on second violation 3. Banned from platform 4. In extreme cases local law enforcement is notified	1. Report 2. Dispute resolution advice 3. Moderation team 4. Restrict 5. Ban	1. User reporting 2. AI monitoring 3. Banned accounts 4. Appeal process for disputes	1. User reports 2. Advice for disputes 3. Warning labels 4. Accounts suspended 5. Blocked by Twitter	1. Report 2. User notified by YouTube 3. Automatic takedown 4. Account suspension 5. Account termination
Content Warnings	For especially graphic and violent photo and video content. Also may have 18+ restrictions on some content.	Text boxes under posts, such as those containing COVID content. Photos and videos containing sensitive or graphic content may appear with a warning to let people know about the content before they view it. This warning appears when viewing a post in feed or on someone's profile.	When messages are reported for inappropriate content, flagged message content will show the following warning: "this message may contain unwanted or harmful content."	Warning labels placed over any tweets or content believed to be misleading, inflammatory or factually incorrect. Pop up message asking if you've read a tweet before you Retweet or Quote Tweet.	YouTube restricts functions on content that are close to the removal line or could be offensive to some viewers by turning off comments, suggested videos and likes. Age restrictions for content over 18+ automatically excludes younger viewers from seeing those videos.

PLATFORM
COMPARISON

Sound Bites

	From Community Standards: “We want people to be able to talk openly about the issues that matter to them, even if some may disagree or find them objectionable”
	From public facing Community page: “It’s our responsibility to foster a safe and supportive community for everyone”
	From top of Professional Community Policies: “We want LinkedIn to reflect the best version of professional life. This is a community where we treat each other with respect and help each other succeed. Be safe... Be trustworthy... Be professional”
	From Help Center: “Twitter is a social broadcast network that enables people and organisations to publicly share brief messages instantly around the world. This brings a variety of people with different voices, ideas, and perspectives. People are allowed to post content, including potentially inflammatory content, as long as they’re not violating the Twitter Rules”
	From Community Guidelines: “When you use YouTube, you join a community of people from all over the world. The guidelines help keep YouTube fun and enjoyable for everyone. If you see content that you think violates these guidelines, use the flagging feature to submit it for review by our YouTube staff”

The trademarks for the platforms referred to above are the property of their owners. Friend MTS is not connected with, and does not endorse, any of those platforms.

Offensive and misleading content that remains may also be downgraded in a platform's algorithm so that it is not distributed as widely, and/or it may be officially 'demonetised' so that ads are not served alongside the content, so to ensure that advertisers are not seen to be supporting such material.

All platforms publish their guidelines and policies in one form or another. Many platforms will have some more user-friendly documents in this domain, as well as a formal outline of all the rules and regulations contained within the company's terms of service. A platform may also publish a regular report summarising all of its activity around the regulation of harmful content.

How easy this information is to find and digest by users varies from platform to platform, and sometimes depending on how the platform is accessed, so that it might be easier to find via a web browser than within an app.

Most platforms will also have statements regarding freedom of expression, usually to the effect that protecting free speech is important to the company and that as a result they might host content that some people consider to be offensive or controversial. They may also talk about the importance of context, in that content might be harmful in some contexts but not others.

All platforms are constantly having to balance the need for freedom of expression and the need to restrict online harms. While some platforms might be accused of often using the former to excuse the latter – or of being inconsistent in the way they tackle this balancing act – it is important to acknowledge that it is a genuinely tricky challenge for everyone.

SECTION 5

COMMERCIAL PRESSURES

As well as navigating existing laws – and any new legal obligations put in place by things like the Online Safety Bill or Digital Services Act – internet platforms face another significant pressure point when it comes to dealing with harmful content – the demands of the advertising industry.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

Many internet platforms rely on advertising sales for some, and in many cases most, of their income. This means that if big brands and/or ad agencies become sufficiently concerned about harmful content on a platform that they stop buying ad space, that is a major problem. Both YouTube and Facebook have been on the receiving end of advertiser boycotts.

Unsurprisingly, most platforms that rely on advertising income are keen to ensure that they overcome any concerns their advertisers may have in this domain. This often involves working with the big ad agency groups and the advertising industry's trade bodies to identify the key concerns and what is required to reassure advertisers that those concerns are being addressed.

That includes the World Federation Of Advertisers, which has been involved in the creation of the Global Alliance For Responsible Media.

In September 2020, a number of the big internet platforms – including Facebook, YouTube and Twitter – agreed to a common framework with the advertising sector to deal with harmful content and the specific concerns of advertisers.

That particular initiative included an agreement that standards should be developed across the industry for how harmful content is defined and reported. There was also a commitment to have "independent oversight on brand safety operations, integrations and reporting" and "to develop and deploy tools to better manage advertising adjacency".

"The issue of harmful content online has become one of the challenges of our generation. As funders of the online ecosystem, advertisers have a critical role to play in driving positive change and we are pleased to have reached agreement with the platforms on an action plan and timeline in order to make the necessary improvements. A safer social media environment will provide huge benefits not just for advertisers and society but also to the platforms themselves"

Stephan Loerke, CEO, World Federation Of Advertisers

© November 2021 Friend MTS Limited. All rights reserved.

"The world is growing increasingly polarised, driven by inequities and disruption, and often that division is powered by harmful content. While such content can be found in any media type, we focus on digital as this is the least regulated environment, with many governments still in the process of defining illegal content. In addition, advertiser transparency and placement controls in user-generated platforms that are open – those with no delay in publishing content – have been underdeveloped. For advertisers who have invested heavily in these platforms, the danger of seeing their brands next to harmful content has become a major issue. No one wants to be inadvertently funding people intent on causing damage to society. GARM is looking to create common cross-industry codes that improve industry transparency and control to effectively contain harmful content and remove advertising from any content that damages brands. Success will be a continual reduction in harmful content on all platforms, underpinned by widely adopted standards, harmonized reporting, independent oversight and common tools"

Global Alliance For Responsible Media

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

In terms of the bigger picture, pressure from the advertising sector can have an impact in two different ways: a narrower impact and a wider impact.

The biggest concern for many brands and ad agencies is that their content is never adjacent to harmful content. Which means the promotional messaging they are paying to post or push does not play before, during or after – or appear alongside – content that is deemed harmful.

In the wider scheme of things, this has a narrower impact. It does require platforms to have policies that identify the kind of harmful content that advertisers are concerned about and systems in place to ensure advertiser messaging does not appear adjacent to that content.

It also means that creators and influencers who have a commercial relationship with an internet platform – ie they are part of a scheme that allows them to share in the income generated by any advertising serviced alongside their content – will be disincentivised from making and uploading any content that is deemed harmful by advertisers and likely to be 'demonetised'.

But it doesn't necessarily mean the harmful content is blocked or removed. By investing in filtering and monitoring systems to safeguard advertisers, platforms are more likely to identify harmful content which their own policies may require to be removed. But plenty of content may be allowed to remain, it just won't get advertising placed next to it.

However, some brands and ad agencies might have concerns beyond the content they directly appear next to. Where a platform is deemed to be failing to deal with harmful content on its platform in general, a brand or ad agency may feel under pressure to withhold advertising for general corporate reputation reasons. In this way, advertisers can have a wider impact.

This is obviously more likely if there are high profile campaigns calling for a boycott – which there sometimes are – or particular controversies that get a lot of mainstream media coverage.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

SECTION 6

NEXT STEPS

The debate around harmful content remains prolific, both within political circles and beyond.

Incidents routinely occur on the big digital platforms that result in widespread calls for new rules and regulations. Though pretty much every proposed reform then proves to be controversial, motivating others to air their criticisms, questioning the effectiveness of the reforms and raising major concerns about privacy and free speech.

The Online Safety Bill in the UK and the Digital Services Act in the EU continue to go through the motions, with amendments aplenty to consider. In the months ahead revised versions of those legislative reforms will come together, with votes to follow.

Once those measures become law, it will be fascinating to see how they work in practice, how the big digital platforms respond, to what extent they really deal with online harms, and whether the concerns of critics regarding privacy and free speech prove to be valid.

It will take some time to truly assess the impact of both the UK and EU approaches. Meanwhile, law-makers in other jurisdictions will make and debate their own legal reforms.

And whatever happens, it seems likely that controversies around harmful content will continue to occur on a regular basis, meaning we can expect these debates in one form or another to keeping going in the long term.

© November 2021 Friend MTS Limited. All rights reserved.
www.friendmts.com

Friend MTS

© 2021 FriendMTS