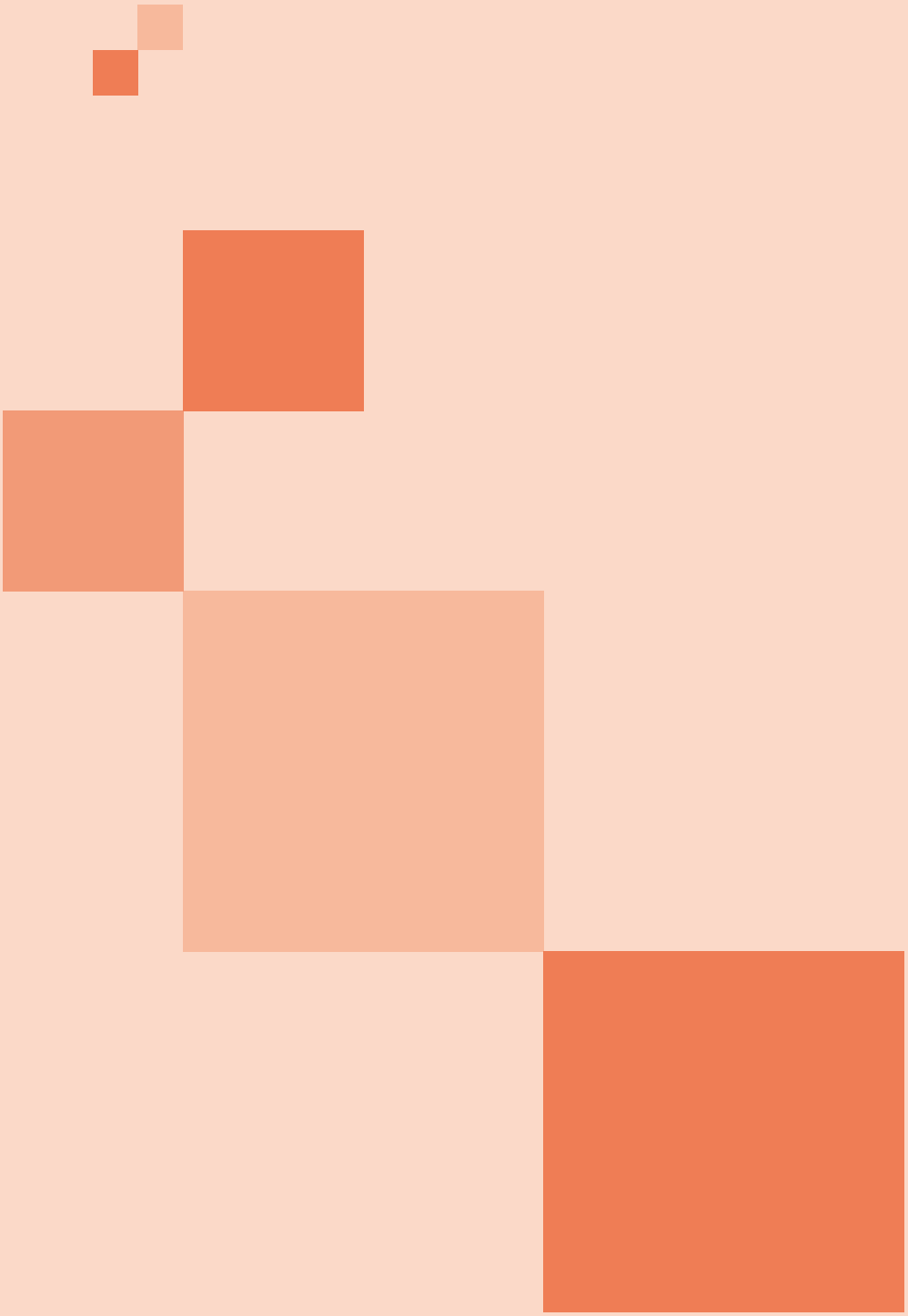
GLOBAL PARTNERS DIGITAL

A Rights-Respecting Model of Online Content Regulation by Platforms

May 2018



Contents

This white paper was authored by
Charles Bradley and Richard Wingfield.

The authors are particularly grateful to all
those who provided comments and feedback
on earlier drafts of this white paper.

Design by Brendon Cleaver.

Published in London, 2018
by Global Partners Digital.

Foreword

Hardly a day goes by without an online platform making the news. Whether in the context of 'fake news' and misinformation, online abuse and harassment, or the use of algorithms to decide what content we see, the activities of platforms are coming under scrutiny as never before.

In many respects, this is unsurprising. These platforms are now a key means of communication for billions of users around the world, who depend on them to build and maintain relationships, exchange information, and express themselves. This gives these platforms significant power and influence, as well as a serious responsibility, guided by international norms, to respect the right to freedom of expression of their users.

In recent years, we have seen growing pressure on platforms from governments to 'do more' to remove unlawful and harmful content – whether by removing it more rapidly, or introducing algorithms and other tools to detect it automatically.

So far, these calls have received a mixed response from platforms, who prefer to deal with these problems on their own or, less frequently, through multistakeholder partnerships. But seemingly arbitrary and non-transparent decisionmaking by platforms has resulted in further criticism, particularly by those concerned by potentially adverse impacts upon freedom of expression.

There is a clear need for a model of content regulation by platforms which is both consistent with their responsibility to respect their users' right to freedom of expression and which addresses the legitimate concerns of governments, making more interventionist proposals unnecessary. This white paper seeks to propose such a model – one which respects human rights and meets the legitimate interest of governments in having unlawful and harmful content removed.

In designing the model, we have kept in mind two particular considerations. First, the need to ensure that a variety of relevant actors and stakeholders are included in framing and developing responses to these issues. Though responses have so far largely come from platforms and governments, neither can solve the challenges alone. A wide range of stakeholders have an interest in addressing these challenges, including civil society organisations, academia and law enforcement agencies. All stages of our model involve their participation.

A second consideration is the diversity of online platforms, and the corresponding need for the model to be adaptable. A platform which allows individuals to share and comment on videos is very different from one which allows users to host files; similarly, a social media platform with billions of users is very different from a start up app run by a small team and with a few thousand users. A wide range of different types of platforms are dealing with the challenges of unlawful and harmful content, and appropriate responses will vary depending on their size, type, business model, and user base. There is no 'one size fits all' and we have designed our model to be adaptable to different platforms based on these factors.

This white paper is primarily addressed to three audiences: to online platforms themselves as a model which they should adopt, to governments as a model that they should support and facilitate through ensuring an appropriate legal and regulatory framework, and to civil society as a model for which they can advocate in their engagement with platforms and governments.

We recognise that these issues are relatively new, rapidly developing, and that there are a wide range of positions and opinions on how they should be addressed. This is why we took the decision to publish our thoughts in a white paper, as a set of proposals which can be debated and discussed. Our model, after all, represents only one possible approach to the questions raised.[1] Some will disagree with certain elements; others may suggest areas for revision or further development. It is also the case that many platforms are already undertaking actions which mirror aspects of our model, which we, of course, commend and welcome. We invite any and all thoughts on this white paper, and hope that it will, in any event, stimulate debate and discussion in a field where progress is sorely needed.

Executive summary

Section 1

Here, we look at the problem which this white paper seeks to address, including the context in which any model of online content regulation must be considered. It also sets out the scope of this white paper – including defining key terms.

Section 2

This section details our proposed model of online content regulation by platforms. The model comprises three stages:

- First, **the development of Terms of Service** by platforms which set out the different forms of unlawful and harmful content which are restricted, taking into account the need for platforms to meet the legitimate needs of their users. These would be sufficiently precise for users to be able to regulate their own conduct and would be developed and reviewed with input from relevant stakeholders.
- Second, **the implementation of these Terms of Service**. This would involve a triaging system whereby flagged content would be passed to specialised teams or individuals based on the particular category it falls into. After judging whether the content should be provisionally removed, the team or individual would determine – based on clear guidelines – whether it is prohibited by the Terms of Service. The final decision would include input from the user who generated or uploaded the content, and, where needed, external expertise.
- Third, **the establishment of a grievance and remedial mechanism**, allowing users to challenge decisions made to remove content (or suspend accounts), and to obtain an effective remedy where successful.

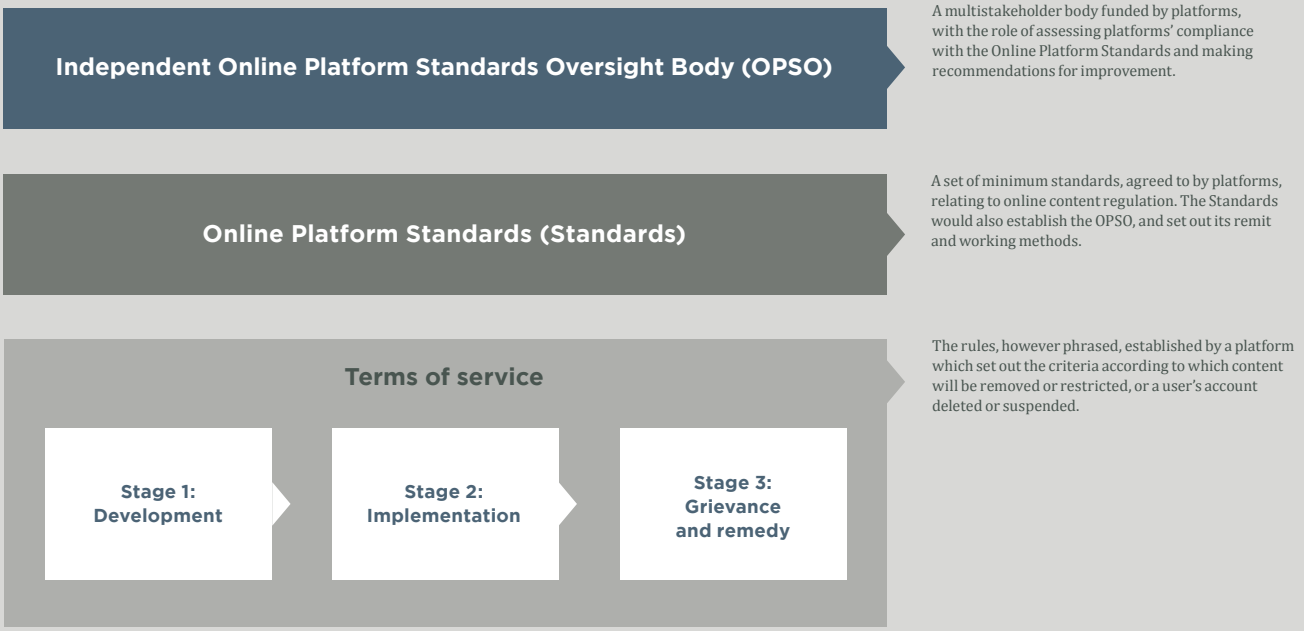
Section 3

Building on the adoption of the model by platforms, we also propose, in Section 3, the establishment of a new global oversight mechanism, the Independent Online Platform Standards Oversight Body (OPSO), funded by platforms, but made up of multistakeholder representatives. The OPSO would assess compliance by platforms with a set of Online Platform Standards (the Standards), developed by independent experts following consultation with the platforms themselves and other stakeholders, and setting out minimum agreed standards for the three stages outlined above. The OPSO would review platforms' compliance with the Standards and publish periodic reports with recommendations on areas for improvement where needed.

Section 4

Finally, in Section 4, we turn to the role and responsibilities of governments. Governments have a critical role to play in ensuring that the legal and regulatory framework which applies to platforms is one which enables platforms to adopt the model which we propose, and avoids imposing restrictions on content or inappropriate intermediary liability. There is also a potential role for governments to take steps to encourage platforms to increase their adherence to the principles which underpin the model.

The whole model



01
Framing the white paper

Before any model of online content regulation can be developed and considered, it is important to identify the specific problem which needs to be addressed and understand the context in which platforms operate.

In this first section, we set out the problem that this white paper seeks to address; namely, the challenges faced by online platforms, users and governments as a result of the ability for unlawful and harmful content to be generated and shared on those platforms. We also look at the context in which this problem manifests. Finally, we set out the scope of this white paper and define a number of key terms used throughout it.

The problem

At its essence, the problem is the current failure of governments and platforms to address the harms resulting from the existence of unlawful and harmful content online. These harms are many, and include:

- **Online content which poses risks to public safety or national security**. For governments, whom this predominantly concerns, the overriding focus is on getting this content removed from an online space which is mainly under the control of the platforms themselves.
- **Online content which causes specific harms to individual users (and others)**. This is of concern to both governments and platforms, as well as users themselves, of course. Harms here include: emotional harm caused to an individual user through hate speech or abuse directed toward them; financial harm caused to holders of copyrighted material which is shared freely and unlawfully online; the sexual abuse suffered by children for the purposes of creating images and videos to be shared online; and the harm caused to individuals who become radicalised online.

These problems are an inevitable consequence of the existence of platforms which facilitate and enable individuals to generate and share content online. But to fully understand the nature of this problem, it is necessary to bear in mind three additional contextual factors.

First, **the scale of online content being generated and shared and its impact upon the exercise of the right to freedom of expression**. About half of the world's population is now connected to the internet,[2] including over 70% of 15-24 year olds.[3] The increase in connectivity has been rapid, trebling over the last decade or so. Connected to this rapid growth in access, the number of users of online platforms, and the amount of content generated and shared on those platforms, have also risen significantly. Facebook, the world's largest social media platform, has more than 2 billion active users each month.[4] As of July 2015, more than 400 hours of video were being uploaded onto YouTube every minute.[5] Every day, hundreds of millions of tweets are sent on Twitter.[6]

In 2016, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression noted that private companies like these "[wield] enormous power over digital space, acting as a gateway for information and an intermediary for expression", adding that the "contemporary exercise of freedom of opinion and expression owes much of its strength" to them.[7] In societies where the offline exercise of the right to freedom of expression is constrained by censorship or state regulation of the media, online platforms may be one of the only ways that individuals are able to exercise that right.

These problems are an inevitable consequence of the existence of platforms which facilitate and enable individuals to generate and share content online.

Second, the development and adoption of the UN Guiding Principles on Business and Human Rights (the Guiding Principles) in 2011 has, for the first time, established **a clear framework for the role of businesses when it comes to human rights**. The Guiding Principles are clear that all businesses have a responsibility (albeit not a legal obligation) to respect human rights, to avoid causing or contributing to adverse human rights impacts through their own activities, and to address such impacts when they occur. As a result of the Guiding Principles, it is now possible to more clearly identify the responsibilities of online platforms when it comes to their users' right to freedom of expression – the steps they should take both to respect that right and avoid adversely impacting upon it, and to address impacts when they occur.

Third, **the role of platforms in relation to content has changed in recent years**. Traditionally, a distinction could be made between platforms which merely hosted content and made no editorial decisions about that content, and publishers which did make such decisions. This distinction is crucial since there exist a number of legal regimes across the world – such as Article 14 of the European Union's Directive on electronic commerce – which exclude liability for content merely hosted by a platform or other company unless they are notified, or otherwise become aware, of content being hosted which is unlawful.[8] As such, platforms which merely host content have no proactive duty to monitor that content.

But online platforms are no longer entirely neutral in hosting and making available content online. Many platforms use algorithms which determine the manner and order in which content is available, make recommendations to users to access certain content, and promote targeted advertising. Many platforms also proactively monitor content to make decisions about its compliance with their Terms of Service. As such, they are no longer passive, neutral hosts of content generated by their users.

Extracts from the UN Guiding Principles on Business and Human Rights

Principle 11 requires businesses to respect human rights. This means that they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved.

Principle 13 sets out the responsibility to respect human rights as including a requirement that businesses avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur.

Principle 14 makes clear that the responsibility of businesses to respect human rights applies to all enterprises regardless of their size, sector, operational context, ownership and structure.

Principle 15 requires businesses to enable the remediation of any adverse human rights impacts they cause or to which they contribute.

And the greater their involvement in making decisions about the content we see, the greater their impact upon users' right to freedom of expression and thus the greater their obligations under the Guiding Principles.

Despite this, and despite the adoption of the Guiding Principles, and collaborative efforts such as the Global Network Initiative (GNI) to ensure ICT businesses respect human rights,[9] concerns persist that **online platforms are failing to respect their users' right to freedom of expression**. Between 2014 and 2016, the Center for Technology and Society of Fundação Getulio Vargas Rio de Janeiro Law School analysed the Terms of Service of 50 major online platforms in order to assess how they dealt with human rights, including the right to freedom of expression.[10] Their conclusion was clear: "*Online platforms offer few guarantees in their policies on preserving the right to freedom of expression. There is a lack of clear and specific information in the Terms of Service on which content is allowed or not in the platform. There is also little commitment to offering users justification, notice and the right to be heard when content is removed by the platforms' own initiative or after notification from third parties*".[11]

As a result of the Guiding Principles, it is now possible to more clearly identify the responsibilities of online platforms when it comes to their users' right to freedom of expression – the steps they should take both to respect that right and avoid adversely impacting upon it, and to address impacts when they occur.

In 2017, the Ranking Digital Rights Corporate Accountability Index reviewed 22 major internet companies and found that they published little information on their policies which affected users' right to freedom of expression, when they removed users' content or suspended their accounts, or what grievance and remedial mechanisms existed for users to challenge decisions to remove content or suspend accounts.[12]

As well as this lack of transparency, there have been a number of high profile examples of inappropriate content removals. In 2017, YouTube deleted a number of videos containing evidence of atrocities in Syria.[13] On Twitter, the accounts of verified news channels and users who have complained of harassment have been suspended.[14] In 2016, Facebook deleted posts of a famous photograph of a napalm victim in the Vietnam War.[15] While, in some instances, platforms have sought to remedy the situation, it has often only been following public pressure. The scale of day-to-day, lower profile instances of inappropriate content regulation is unknown, partly as a result of the lack of any meaningful transparency about moderation decisions from the online platforms themselves. This lack of transparency also reinforces the difficulty of ensuring awareness of when and why mistakes have been made.

There is a final point to consider, which makes the need for a rights-respecting response more critical. National governments are increasingly responding to the problems outlined above with **legislative and policy proposals which are likely to lead to greater regulation of online content and the imposition of increased liability on platforms**.

These proposals focus largely on simply getting content removed more quickly, rather than addressing the underlying causes of the social problems manifesting online. The Network Enforcement Act (or NetzDG) in Germany is perhaps the most high-profile example, requiring online platforms with more than two million subscribers to remove "manifestly unlawful" content within 24 hours, and imposing fines of up to €50 million for non-compliance. At the EU level, the Home Affairs Commissioner has demanded that platforms take down illegal content within two hours and suggested that the Commission would introduce legislation if platforms failed to do so voluntarily.[16] In the UK, a Home Office Minister proposed taxing online platforms which fail to take down 'radical' or 'extremist' content,[17] and the Prime Minister has called for platforms to take down such content within two hours.[18] In France, President Macron promised to introduce a new law to prohibit 'fake news' during elections.[19] These proposals would have serious consequences for freedom of expression online, incentivising online platforms to take down lawful content rather than risk fines or other sanctions.

Taken together, these factors reveal a landscape in which the right to freedom of expression is increasingly being exercised online, with online platforms having a clear responsibility to respect that right as exercised by their users. However, that right is not being fully respected in practice, and is further threatened by legislative and policy proposals from governments concerned that online platforms are failing to deal with the challenges of unlawful and harmful content sufficiently seriously.

The scale of day-to-day, lower profile instances of inappropriate content regulation is unknown, partly as a result of the lack of any meaningful transparency about moderation decisions from the online platforms themselves.

The scope of this white paper

The title of this white paper – 'A Rights-Respecting Model of Online Content Regulation by Platforms' – expresses a clear, straightforward purpose: to propose a model of online content regulation by platforms which is consistent with international human rights law and standards, primarily the right to freedom of expression. However, to make clear the precise scope of this white paper – what it includes and doesn't – some of these terms require definition. Below, we set out what we mean by 'online content', 'regulation' and 'platforms', as well as certain other terms which we use throughout the white paper.

'Online content': As noted above, this white paper proposes a model for regulating online content by platforms which is consistent with international human rights law and standards, primarily the right to freedom of expression. Our focus is therefore on those forms of online content protected by that right. 'Content' and 'expression' are not, however, synonymous, and some consideration should be given to what each term means and what differences, if any, exist between them.

The primary source of the right to freedom of expression in international human rights law is Article 19 of the International Covenant on Civil and Political Rights (ICCPR).[20] Neither Article 19 nor the UN Human Rights Committee's interpretation of that right in its General Comment No. 34 provides a full definition of the term 'expression'. Instead, both give a partial definition.

Article 19 provides that 'expression' includes "information and ideas of all kinds" in whatever form, and General Comment No. 34 notes that it includes "communications of every form of idea opinion capable of transmission to others".[21] Although Article 19 was drafted prior to the advent of the internet, General Comment No. 34 also makes clear that the means of expression falling within the scope of Article 19 include "electronic and internet-based modes of expression".[22] Although these are not fully exhaustive definitions, they are extremely broad, and 'expression' can therefore be said to encompass the transmission of information, ideas and opinions of all kinds and in whatever form.

The term 'content', despite being widely used, does not have a universally agreed definition. The Oxford English Dictionary provides a definition of "information made available by a website or other electronic medium".

However, the ubiquity of the term suggests that anything that is made available online, whether read, seen or heard, can be considered 'content'. We do not propose to provide any definition of the term 'content' in this white paper, and instead use it on the basis that it includes anything which can be read, seen or heard online.

The broad scope of both 'content' and 'expression' makes clear that there is a large degree of overlap and suggests that there is no obvious form of online content which could not be considered as a form of 'expression'. This conclusion is reinforced by a comparison of online forms of content with traditional offline forms of transmission of information, ideas and opinions. The internet and other ICTs allow for a wide and increasing range of forms of content which can be broken down into three categories: (i) those which replicate, to a large extent, traditional offline forms of transmission of information; (ii) those which are identical to particular offline forms of transmission of information save that they exist in a purely digital, rather than a physical, format; and (iii) new forms of transmission of information which exist solely in the digital environment with no obvious offline equivalent.

	Offline	Online
Category 1: Online information replicating offline information	Letters Physical documents Newspapers Speech	Emails Digital documents Online newspapers Social media posts
Category 2: Online and offline information identical, save for format	Photographs Video Audio	
Category 3: New forms of online information with no obvious offline equivalent		Hyperlinks Emojis and animojis Gifs

Given the equivalence between their offline and online forms, the forms of transmission of information which fall within categories 1 and 2 can easily be considered as content which is protected by the right to freedom of expression. Although, in the case of online forms in category 3, there has been no authoritative determination of whether they are protected forms of 'expression', the broad scope of both terms means that, for the purposes of this white paper, we have proceeded on the basis that they should be considered as protected forms of expression.

We therefore use the term 'online content' to refer to any information, idea or opinion available online, in whatever form and whether intended to be read, seen or heard; and, further, we consider that all online content falls within the scope of the right to freedom of expression.

'Regulation': The term 'regulation' can include a range of measures taken by a variety of state and non-state actors, including platforms themselves, which result in limitations on the availability of certain content or the ability to express or communicate certain information, ideas or opinions. For the purposes of this white paper, we only look at regulation undertaken by the platforms themselves. This includes moderation of content based on a platform's Terms of Service following flagging by users, state actors and others. It would not include any action taken by platforms in response to court orders or other demands by state actors, which we consider to be a distinct issue.

In this white paper, we do not look at this related but distinct issue of platforms' responses to demands or requests from governments, courts or other state actors for the removal of content which is in breach of national law, as opposed to the platforms' Terms of Service. We recognise the importance of this issue from a freedom of expression perspective, but focus on platforms' Terms of Service in this white paper for two main reasons:

First, because in Section 4 we call on governments to ensure that their national legal frameworks do not contain restrictions applying to online content which are inconsistent with their obligations with respect to the right to freedom of expression. If both national legal frameworks and platforms' Terms of Service are consistent with the requirements of the right to freedom of expression under international human rights law, there would be few cases where content was prohibited by national law and not also prohibited by a platform's Terms of Service.

Second, although figures are limited, it appears to be the case that the vast majority of content removal by platforms is being done on the basis of its Terms of Service rather than following a request by a state actor. YouTube – one of the only major platforms to provide transparency on this issue – records that in 2017, between January and June, it removed around 20,000 videos following requests from the government, the police, a court or other state actor. [23] Between October and December 2017, just half that time, it took down over 8 million videos as breaching their Community Guidelines.[24]

We use the following terms, all of which are examples (though not exhaustive) of 'regulation' by platforms:

- **Removal of content**: making inaccessible online content which was previously accessible (sometimes referred to as a 'takedown' of content);
- **Moderation of content**: reviewing content in order to make a determination as to whether it should be removed;
- **Restriction of content**: preventing a certain form of content from being made accessible online at all;
- **Deleting (or suspending) an account**: permanently (or temporarily) preventing a user from being able to make content accessible online, or otherwise use an account.

'Platform': As with the term 'content', the term 'platform' is commonly used, but with no universally agreed definition. In its broadest sense, a 'platform' is any hardware or software which is used to host an application or service, or any form of technology on which other technologies are built. As with the term 'content', we do not propose to provide an exhaustive definition of the term 'online platform', instead preferring to use the term to broadly include any software-based facility which enables users to generate, host or access content online. In practice, the extent to which the issues we have identified will be of relevance to a particular platform will vary greatly. For example, while content regulation will be highly relevant to social media and search platforms, they will be less so to a video-on-demand platform.

'Terms of Service': By 'Terms of Service', we refer to any rules (regardless of phrasing or format) established by a platform which set out the criteria according to which content will be removed or restricted, or a user's account deleted or suspended. This includes 'Community Standards', 'Participation Guidelines', 'Rules', etc.

'Unlawful or harmful': This white paper looks at a model of online content regulation that could be adopted by platforms globally. As such, when we use the terms 'unlawful' and 'harmful' to categorise content that platforms could (or even should) regulate, we are not referring to content which is unlawful or harmful by any national standards, but by global standards.

'Unlawful' therefore refers to content which is prohibited by international human rights law (such as propaganda for war; advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence; and images of child sexual abuse). 'Harmful' refers to content which could be permissibly restricted by virtue of Article 19(3) of the ICCPR, i.e. its restriction is necessary for respect of the rights or reputations of others; or for the protection of national security or of public order, or of public health or morals.

02
A model for rights-respecting content regulation

In this section we set out our model for how platforms can regulate content in a human rights-respecting manner through their Terms of Service. The model we propose can be divided into three stages: the development of Terms of Service, their implementation, and the provision of a grievance and remedial mechanism. It is the case that many online platforms will already be compliant with some aspects of the model; however, no platform at the time of writing is fully compliant with the model as a whole.

Section 2.1. Developing Terms of Service

Overview

- *Platforms should develop, and periodically review, Terms of Service which comprehensively set out the forms of content which are restricted.*
- *The Terms of Service should be made easily available and accessible for users.*
- *The Terms of Service should be sufficiently precise so that users can regulate their conduct.*
- *The Terms of Service should categorise the different forms of restricted content, supplementing this with more detailed interpretation and guidance.*
- *The development and review of Terms of Service should involve consultation and engagement with a range of relevant stakeholders.*

A platform's Terms of Service can serve a number of purposes. They can help project and elucidate a platform's values. They can instil confidence in potential users that using the platform will be a safe experience. They can, when they form part of a contract between the platform and the user, provide a legal basis for the platform to take action against a user.

Why Terms of Service are needed

From a human rights perspective, Terms of Service serve two particular purposes. First, they make clear what forms of content the platform will remove or restrict, allowing for comparison with the justified limitations on freedom of expression under international human rights law. Second, they enable users to know, with a reasonable degree of confidence, under what circumstances content they wish to make available will be removed or restricted, ensuring transparency and certainty.

Terms of Service may also include other aspects of a platform's operations, or its relationship with its users and third parties. They may be titled as 'Community Standards', 'Community Guidelines', 'Content Policy' or something else. Here, we use 'Terms of Service' as a catch all, referring to the platform's rules relating to content.

We believe that the development of Terms of Service is not just beneficial, but a responsibility of platforms under international human rights law and the Guiding Principles. As is noted in Section 1, the right to freedom of expression includes online expression as well as offline expression.[25] (Indeed, all human rights apply online as well as offline).[26] Principle 11 of the Guiding Principles provides that "business enterprises should respect human rights" and that this means that "they should avoid infringing on the human rights of others". We believe that, taken together, these two principles mean that platforms – in order to ensure a consistent degree of protection of human rights – have a responsibility not to restrict freedom of expression exercised via their platforms in a way which is inconsistent with international human rights law and standards.

Under international human rights law, restrictions on freedom of expression are only permissible when they are "provided by law" (to use the wording in Article 19). This means that any restriction must be "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly and it must be made accessible to the public".[27] While Article 19 was drafted to set out the obligations of states, we believe that the responsibility of businesses to respect human rights is best met through applying the same principles, as far as possible. As such, we believe that platforms should not restrict freedom of expression unless the restrictions are "made accessible to the public" and "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly". This, in essence, is what Terms of Service should do.

Availability and accessibility

Terms of Service should be easily accessible for users both during use of the platform and, where registration is required, at the point at which the user signs up to the platform. While it is, of course, up to the user to decide whether and when to review a platform's Terms of Service, the platform should take reasonable steps to make users aware of their existence. They should not be contained in a long, dense user agreement; nor should they be difficult to find on the platform's website. Instead, they should be published as a self-contained resource, and be quickly and easily accessible on the platform's website. In addition, the Terms of Service should, as far as possible, be in plain language and accessible formats, and available in the languages that their users understand. Where they are revised, users should be notified in advance of the changes being made.

Sufficient precision

Because of the need under Article 19 for "sufficient precision" when restricting freedom of expression, only setting out the types of content that will be moderated in any Terms of Service would not be enough to meet the requirements of the first criterion for permissible restrictions under Article 19. States, for example, would meet this obligation through specific legal provisions of general applicability, accompanied by some form of elaboration (e.g. explanatory notes published alongside legislation, guidance from relevant government departments, or guidance from the police or prosecution authorities). Interpretation of terms by courts can also help provide clarity on the circumstances when particular forms of expression will be prohibited. We believe that platforms should provide an equivalent degree of clarity so that users are able to regulate their conduct (i.e. the content they upload, generate and seek to access) accordingly. This means that as well as developing Terms of Service, platforms should ensure that they provide sufficient detail – whether through accompanying documents or in the Terms of Service themselves – to enable users to know, with a reasonable degree of certainty, whether particular content is or is not restricted. The level of detail provided by platforms' existing Terms of Service currently varies greatly.

Example 1: Facebook and 'graphic violence'

Facebook's Community Standards (as of March 2018) explained that: "*Facebook has long been a place where people share their experiences and raise awareness about important issues. Sometimes, those experiences and issues involve violence and graphic images of public interest or concern, such as human rights abuses or acts of terrorism. In many instances, when people share this type of content, they are condemning it or raising awareness about it. We remove graphic images when they are shared for sadistic pleasure or to celebrate or glorify violence".*

It is unclear – in the absence of clarification or examples – what 'graphic images' or 'graphic violence' actually mean in practice. Two simple examples illustrate the problem. If a photograph of a seriously injured child is posted with no comment, it is not clear whether this would be removed or left up. Nor is it clear how Facebook would determine whether the photograph was being shared for 'sadistic pleasure' or to evidence a human rights abuse.

Take another hypothetical example – someone sharing a video of a member of an ethnic minority being severely beaten by the police. Would it be taken down if it was accompanied by a smiley face emoji (on the basis that the user was celebrating violence) but left up if it was accompanied by an angry face emoji (on the basis that the user was raising awareness of legitimate concerns relating to police brutality)?

The Terms of Service provided no indication as to whether or not these forms of content would be removed, nor how such decisions would be made.

Example 2: Twitter and 'hateful conduct'

'Hateful conduct' is a broad term; however, Twitter's 'hateful conduct policy' (as of March 2018) provides a definition of what they mean by 'hateful conduct': *"[to] promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease".*

This definition is accompanied by a non-exhaustive list of examples:

- violent threats;
- wishes for the physical harm, death, or disease of individuals or groups;
- references to mass murder, violent events, or specific means of violence in which/with which such groups have been the primary targets or victims;
- behaviour that incites fear about a protected group;
- repeated and/or or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone.

While the examples are non-exhaustive, they give a clear indication of the sorts of content that will be considered as 'hateful' and which will be removed.

In practice, the Terms of Service which platforms have so far developed tend to set out broad categories of the different forms of unlawful or harmful content which they prohibit; for example, 'hate speech' or 'graphic violence'. As noted in Section 1, it is important to look at each issue being addressed independently, and to respond in an appropriate fashion, so we support the categorisation of forms of unlawful and harmful content. We detail possible categories later on in this section, and, in Section 2.2, propose a triaging procedure for platforms when responding to content which has been flagged, using these categories to help determine how to respond. However, regardless of which broad categories of restricted content are used, there are a range of ways that this "sufficient precision" criterion can be met:

- Platforms could simply provide more detailed interpretation or guidance in the Terms of Service themselves.
- If platforms have concerns that this would make the Terms of Service too long or complex, they could retain broad, simple categories in the Terms of Service with more detailed interpretation or guidance available via a link.
- Platforms could also provide examples, either hypothetical or based on real instances, of content that would or would not be restricted under each category.

Categorisation of the forms of restricted content

As well as a requirement that any restrictions on freedom of expression be "provided by law", Article 19 of the ICCPR also requires that they be for one of a number of specified purposes, namely (a) for respect of the rights or reputations of others, or (b) for the protection of national security or of public order, or of public health or morals (the permissible limitations set down in Article 19(3)).

International human rights law also requires the prohibition of certain forms of expression: Article 20 of the ICCPR prohibits propaganda for war and advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence. The Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography prohibits, among other things, images of child sexual abuse.

We believe this has two key implications for platforms:

- First, they should restrict content which constitutes propaganda for war; advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence; or child sexual abuse.
- Second, if they are to restrict any further forms of content, such restrictions should be necessary and in pursuance of one of the legitimate aims set out in Article 19(3), i.e. to ensure respect for the rights or reputations of others, or for the protection of national security, public order, public health or public morals.

While none of these forms of expression in the first group are defined within the relevant treaties themselves, sources of interpretation and guidance exist. The 2011 report of the former UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue, for example, provides guidance on the interpretation of these and other forms of expression which are prohibited under international human rights law.[28]

In relation to the second group, while these legitimate purposes are broadly worded in Article 19(3), there are also sources of interpretation and guidance as to how they apply to different types of expression. General Comment No. 34, for example, provides further interpretation and clarification of each of the legitimate aims, and they have also been considered in the jurisprudence of cases brought to the Human Rights Committee on the basis of a violation of Article 19. The General Comments and Recommendations of other UN Treaty Bodies, as well as decisions of other regional and national courts interpreting equivalent provisions protecting the right to freedom of expression, are also illustrative.

The nine categories below are typical of the most common forms of restricted content contained within major platforms' existing Terms of Service. All would, fully or partially, correspond to one or more of the legitimate aims in Article 19(3).

Category of content	Purpose
Threats or incitement of violence	The rights or reputations of others
Facilitating other criminal activity	The rights or reputations of others; protection of public order
The glorification of, or support for, terrorism or organised criminal activity	Protection of national security; protection of public order
Bullying or harassment of other users which does not amount to a criminal offence	The rights or reputations of others
Hate speech against particular groups	The rights or reputations of others
Child sexual abuse	The rights or reputations of others
Adult sexual content	The rights or reputations of others; protection of public morals
Violence and other graphic content	Protection of public morals
Copyrighted and trademarked material	The rights or reputations of others

If platforms propose to restrict content which does not fall into these categories, we believe that they should only do so if it would be consistent within one of the legitimate aims set out in Article 19(3). However, even some of these categories, as a result of their breadth, potentially include both content which is and is not unlawful or harmful. For example, 'adult sexual content' could include pornographic videos which a platform could legitimately restrict, but also images of naked adults, or genitalia, which have an artistic or scientific basis, and ought not to be restricted. As such, it is important that the interpretation or guidance which accompanies the Terms of Service makes it clear that content which does not fall into the exceptions set out in Article 19(3) will not be restricted, even where it falls within the broad category of content.

We recognise that there may be situations where platforms have been (or may be) developed for a specific purpose, or for a particular community, which needs restrictions on certain content to ensure that the platform can meet the legitimate needs of its users. For example, a platform which is developed exclusively for children may want to restrict mildly violent or graphic content which a platform developed for adults would not. Or a platform developed to provide a safe space for a particular minority group, or a vulnerable or marginalised community – such as LGBT individuals or those with mental health problems – may wish to restrict content which, while not offensive, indicates opposition to LGBT rights, or which could trigger anxiety or panic among those with a particular mental health condition.

In such circumstances, we consider that such restrictions would fall within the legitimate aim of 'the rights of others'; with 'others', in this case, referring to the users for whom the platform was designed. Where, however, a platform considers that its specific purpose, or the community that it has been developed for, justifies particular restrictions on content, it should ensure that any such restrictions are both "necessary" and as narrowly drawn as possible while still meeting their users' legitimate needs.

Multistakeholder engagement in development and review

There are a number of benefits for platforms that can be derived from consulting and engaging with a broad range of relevant stakeholders during the development of the Terms of Service. This engagement can bring expertise to the process, and boost confidence in, and the legitimacy of, the Terms of Service which are ultimately developed.

Given the generally global application of a platform's Terms of Service, it is even more important that relevant expertise on particular issues be harnessed to ensure that the final Terms of Service are fit for purpose. Users come from a wide range of backgrounds – linguistic, religious, cultural, and other – which means that a platform is unlikely to have all of the necessary expertise to be able to develop Terms of Service which can apply globally and fairly.

Platforms should therefore engage with all relevant stakeholders and representative and interest groups in developing their Terms and Service and accompanying interpretation and guidance.

The precise stakeholders and groups with which a platform should engage will vary depending on the particular form of unlawful or harmful content which is being considered but may include:

- Experts in freedom of expression generally (such as academics or human rights organisations);
- Groups advocating on behalf of particular vulnerable or marginalised groups, such as women, children, persons with disabilities, LGBTI individuals, ethnic and religious groups;
- Law enforcement agencies;
- Experts in terrorism and radicalisation;
- Linguistic experts;
- Psychologists.

For example, developing Terms of Service and accompanying interpretation and guidance on what constitutes child sexual abuse may require consulting experts on international law (particularly the Convention on the Rights of the Child and its Protocols), children's rights groups, and international or national law enforcement agencies.

Terms of Service and accompanying interpretation and guidance should be periodically reviewed to ensure that they remain fit for purpose, and revised and updated as necessary.

Section 2.2. Implementing Terms of Service

Overview

- *Platforms should ensure that they have the functionality to allow users to easily notify them of content which they consider to be in breach of its Terms of Service (flagging).*
- *Flagged content should then undergo a triaging procedure to determine which category of restricted content it falls most closely under, as well as to filter out content which is manifestly and unambiguously not in breach of the platform's Terms of Service.*
- *The user who generated or shared the content should be informed that the content has been flagged, provided with the reasons why, and given a sufficient period of time to provide any information justifying why the content should not be taken down.*
- *If there is seen to be a risk of immediate and irreversible harm were the content to remain available after being flagged, the content should be provisionally removed pending the outcome of the determination process.*
- *Determination should be made within a reasonable period of time as to whether the content is in breach of the platform's Terms of Service.*
- *Platforms should ensure that sufficient resources are provided to the teams and individuals making determinations, from training and support for staff to the provision of external expertise where necessary. Processes for quality assurance of moderation decisions should also be introduced.*
- *The outcome of the determination should be communicated to both the user who flagged the content and the user who uploaded or generated the content, along with reasons and, where relevant, details of the available grievance mechanism.*

A note on pre-emptive and proactive restriction and removal of content

The model we propose for implementing Terms of Service is one to be used only after content has been published and brought to the attention of the platform as potentially in breach of its Terms of Service. There are calls, particularly from governments, for platforms to restrict content from being made available even before it is published ('pre-emptive moderation') and to proactively monitor content on the platform ('proactive moderation'). Some platforms already undertake either or both of these.

With regards to pre-emptive moderation of content, we recognise that there may be certain very limited circumstances where decisions to moderate content prior to publication could be made by a platform consistently with international human rights law and standards. However, these are limited to those where (i) specific content has already been identified by a human as unambiguously and, regardless of context, in breach of international human rights law (and therefore also the platform's Terms of Service if our model is followed), such as images or videos of child sexual abuse, and (ii) it is a copy of such content that a user has sought to share.

Where automatic processes are able to identify content which is a copy of content a platform has already decided should not be published, it is logical for that process to prevent its further publication. There are examples of this process taking place already, such as in the UK, where the Internet Watch Foundation has developed an Image Hash List comprising hundreds of thousands of hashes of images of child sexual abuse. This hash list is updated daily and distributed to companies who pay for the service.

These companies are then able to use these hashes both to identify images of child sexual abuse which have already been uploaded, and to prevent them from being further uploaded at all.

While an example of best practice, the use of such a process is limited to circumstances where the content is a copy of already identified content, and that content is unambiguously in breach of international human rights law (and so the Terms of Service), regardless of context or other factors. Its utility does not extend to the moderation of content which is new, where the content is not clearly unlawful or harmful, or where context is a relevant consideration. While such a model could therefore potentially play a part in preventing, for example, the publication of copyrighted material in certain circumstances, it is difficult to conceive of other forms of content where it could play a role.

As such, and subject to those certain, limited exceptions, we do not consider that platforms should moderate content prior to publication. As well as the risks to freedom of expression given the absence of the safeguards attached to the model proposed in this section, there are also reasons of practicality. The sheer volume of content which is uploaded for publication makes it almost impossible for it all to be pre-emptively moderated by a platform. The number of people and amount of time required would far exceed the capacity of even the most well-resourced platforms, and would entirely undermine the instantaneous nature of content uploading and sharing.

As it is only ever a small proportion of content which is unlawful or harmful, we believe it is preferable for platforms to focus their resources on content which has been flagged as such, rather than to monitor all content prior to publication.

With respect to proactive moderation of content, the same considerations of scale and practicality apply. However, we note that many platforms are already proactively moderating content, often through the use of algorithms and automated processes. Between October and December 2017, for example, YouTube removed over 6.6 million videos identified as in breach of its Community Guidelines following an automated flagging process.[29] Where platforms do proactively moderate content, the same stages set out below should be followed once content has been flagged as a result of that internal review.

The number of people and amount of time required for pre-emptive moderation would far exceed the capacity of even the most well-resourced platforms, and would entirely undermine the instantaneous nature of content uploading and sharing.

Flagging content

Regardless of any proactive moderation of content, platforms should ensure that they have the functionality to allow users to be able to notify the platform, in a simple and straightforward way, of content which they consider to be in breach of the platform's Terms of Service (flagging), thereby instigating the content moderation process.

For the implementation of the Terms of Service to be effective, including from the perspective of the user who published the content, it is important that sufficient information be provided so that the platform can make an informed determination of whether the content is in breach of its Terms of Service. As such, the platform should require users, when flagging content, to provide the reasons why they consider that it is in breach of the platform's Terms of Service.

Some platforms use a system of 'trusted flaggers', 'superflaggers' or some other mechanism by which individuals or organisations can flag multiple items of content as a result of their particular expertise or historic accuracy in potentially identifying content which is in breach of Terms of Service. If a platform decided to use such a system (and with the important qualification that such systems are not without their critics),[31] this would not negate the requirement for a final human determination.

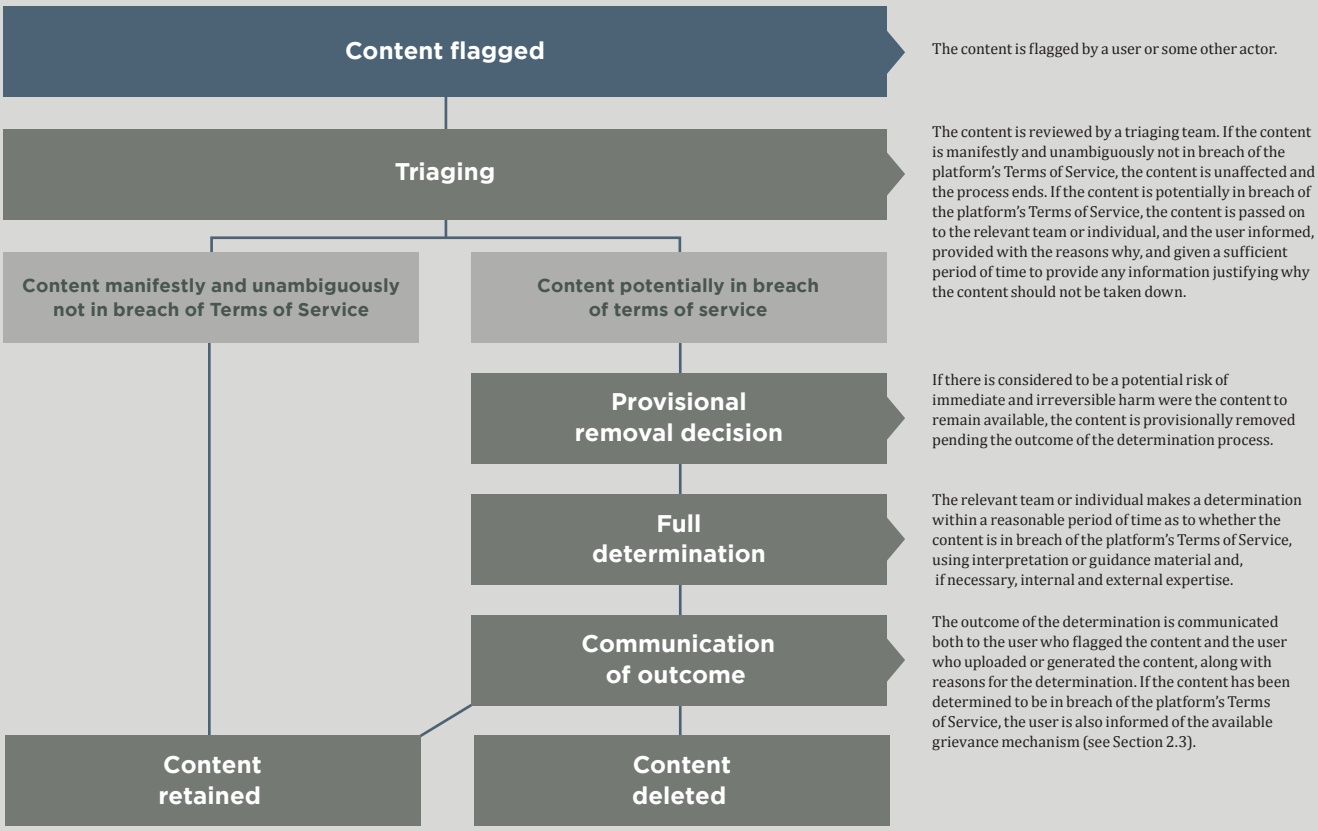
Triaging

Given the wide range of forms of unlawful and harmful content that exist, and the different expertise and stakeholder engagement needed to make determinations, we propose that platforms designate distinct teams or individuals to deal with the different forms of content, using the categories developed under the Terms of Service.

We would also propose that content which has been flagged should undergo a simple triaging procedure to determine which particular category the content falls most closely under, at which point the relevant team or individual will be tasked with undertaking the determination process. It may be the case that this triaging procedure is also able to identify content which is manifestly and unambiguously not in breach of the platform's Terms of Service, in which case the user who flagged the content would be informed that this is the case, and the process would cease.

Unless the content has been identified as manifestly and unambiguously not in breach of the platform's Terms of Service, then (at the same time that the content is undergoing the triaging procedure) the user who uploaded or generated the content should be informed that the content has been flagged, and the reasons why. That user should be given a sufficient period of time to provide any information justifying why the content should not be taken down.

Implementing terms of service



Provisional removal of content

There may be circumstances where it is appropriate for content to be provisionally removed pending the outcome of the determination process. This might apply, for example, in cases where there is a potential risk of immediate and irreversible harm were the content to remain available. In such cases, the user who generated or hosted the content should be informed. Where there is no such risk, such as where the reasons for flagging relate to copyrighted work, content should not be removed until a final determination has been made.

Determination

Once the content has been passed on to the relevant team or individual, a determination should then be made within a reasonable period of time as to whether the content is in breach of the platform's Terms of Service. The team or individuals should use the interpretation or guidance material developed alongside the Terms of Service. There are three additional further considerations that should be taken when platforms develop this procedure:

- First, the platform should ensure that sufficient resources are provided to the teams and individuals making determinations, both in terms of the number of moderators and the amount of time available for moderators to make determinations.
- Second, all staff engaged in content moderation should be given sufficient training and support in their roles. This includes not only introductory training on international human rights law and standards, and their relationship to the platform's Terms of Service, but ongoing and additional training and support where needed. The fact that content which is flagged may be disturbing – such as child sexual abuse imagery or graphic violence – means that the welfare needs of the individuals involved must be considered. Platforms should ensure that they have a rigorous recruitment process in place to ensure that the moderators recruited have the psychological and emotional capacity to undertake the work of moderating such forms of content, and provide the necessary support to moderators. This support could include shorter working hours, regular breaks, and periodic psychological and counselling sessions.
- Third, there may be circumstances where moderators need external support in order to make a decision. This could be as a result of further information and expertise being needed on linguistic, religious or cultural issues. In such circumstances, moderators should be able to – and encouraged to – seek such external expertise, with the same groups identified in Section 2.1 as relevant to developing particular categories of restricted content within the Terms of Service.

Quality assurance

Platforms should introduce processes for the quality assurance of moderation decisions. This might mean inviting 'second opinions' on a selection of decisions to ensure accuracy and consistency; reviews of moderators' decisions and the proportion that are overturned after a second opinion or after an appeal; external review by the groups identified in Section 2.1 of decisions that are made by moderators; or using 'mystery shoppers' to test the moderation procedure from a user's perspective.

Communication of determination

The outcome of the determination should be communicated both to the user who flagged the content and the user who uploaded or generated the content, along with reasons for the determination and – if the content has been determined to be in breach of the platform's Terms of Service – the available grievance mechanism, which we look at in the next Section (2.3).

The role of algorithms and automated processes

One example where automated processes have shown to be successful is the use of hashes by the Internet Watch Foundation in the UK, as detailed above. As well as the clear and objectively unlawful and harmful nature of the content, it is important to note that there is still human oversight of the process, in that analysts check each child sexual abuse image before hashing it and adding it to the Image Hash List. As such, the automated process only kicks in after a particular image has been reviewed by a human, and only applies to that image and copies of it.

Outside of this narrow field, however, the benefits of algorithms and automation are, at least at present, less well established. Indeed, there is clear evidence of the limitations that currently exist in using automation and algorithmic filtering to regulate content. In its recent report, 'Mixed Messages: The Limits of Automated Social Content Analysis',[30] the Centre for Democracy & Technology highlighted a number of substantive limitations to these automated processes in the context of social media platforms. These included:

- The varying levels of reliability in identifying harmful content given significant differences in language use across different platforms, by different demographic groups and depending on the topic of conversation.
- The risk of decisions based on automated social media content analysis further marginalising and disproportionately censoring minority groups and those that face disadvantage.
- The lack of any clear, well-established definitions of forms of harmful content, such as 'hate speech', 'extremist material' or 'radicalisation', which are necessary for effective automated content analysis.
- Differences between what the coders of the tools themselves considered as falling into the categories, often as a result of different cultural backgrounds and personal sensibilities.
- The inability of tools to take into account context – such as tone, the speaker, the audience and the forum – to any meaningful extent. They struggle, for example, to understand jokes, sarcasm, irony and nuance.

These limitations mean that any use of algorithms and automation to filter or otherwise moderate content should be considered very carefully. Although it is understandable that platforms are looking to algorithms and automation to deal with the scale of online content, there are real risks that perfectly lawful and legitimate content may be taken down, and that such moderation will disproportionately impact minority groups and those that already face disadvantage.

As a result any use of algorithms and automation must be accompanied by strong safeguards to mitigate these risks. In particular, we consider that three key safeguards are essential:

- First, there should always be some human oversight of any decisions made by algorithms and automation. While, of course, humans will have developed the processes and authorised their use, we believe that the results of those processes should also be reviewed by a human who will be able to act as a filter against potential removals of content which would breach the right to freedom of expression or disproportionately affect particular groups vulnerable to discrimination.
- Second, to support the procedural requirements of restrictions on the right to freedom of expression, platforms should clearly and transparently publish meaningful and easily understandable information on what processes are being used, for which purposes, and how decisions are made by those processes. This information should be available in the languages used by the users of those platforms as well as in accessible formats.
- Third, the algorithms and automation, and their results, should be regularly reviewed, and the processes refined, to mitigate against the risks identified above.

Section 2.3. Grievance and remedial mechanism

Overview

- *Platforms should establish a grievance mechanism by which users can challenge decisions made to remove content which they have generated or shared, and obtain an effective remedy if they are successful.*
- *The mechanism should comply with the criteria set out in Principle 31 of the UN Guiding Principles, i.e. it should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, a source of continuous learning and based on engagement and dialogue.*

Grievance and remedial mechanisms

However well developed and implemented a platform's Terms of Service may be, mistaken or inappropriate removal of content is inevitable. Such mistaken or inappropriate removals may, however, constitute an adverse impact on the user's right to freedom of expression. The UN Guiding Principles on Business and Human Rights (the Guiding Principles) address this situation, with Principle 22 making clear that where a business identifies that they have caused or contributed to an adverse impact, they should provide for or cooperate in their remediation through a legitimate process. This responsibility reflects the well-established principle in international human rights law that those who have suffered a human rights violation are entitled to an 'effective remedy'.[32]

The UN Guiding Principles also set out in some detail how such a remedy should be provided. Principle 29 states that businesses should "establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted". Principle 31 goes on to set out a number of criteria for a grievance mechanism to be effective.

In the context of content regulation, platforms should establish a grievance mechanism which (i) requires the user to be informed that the content has been removed (or that the platform proposes to remove that content, or that their account has been suspended, as the case may be), (ii) provides an opportunity for the user to challenge that decision, and (iii) provides an effective remedy where the challenge is successful. We further believe that such a grievance mechanism can meet these requirements by fully complying with the criteria set out in Principle 31 of the Guiding Principles.

Finally, under no circumstances should a platform's grievance mechanism exclude the possibility for a user to use alternative state-based grievance mechanisms, such as judicial processes or complaints to a national ombudsman.

However well developed and implemented a platform's Terms of Service may be, mistaken or inappropriate removal of content is inevitable.

Compatibility with Guiding Principle 31

1. Legitimate

The principle of legitimacy requires that the stakeholder groups impacted have trust in the process, and that there is accountability for its fair conduct.

What does this mean for platforms?
Platforms should involve relevant stakeholders in both the design of the grievance mechanism and – where appropriate – in its implementation; for example, by involving the groups identified in Section 2.1 in reviewing decisions that have been made by moderators and appealed.

2. Accessible

The principle of accessibility requires that the grievance mechanism is known to the stakeholders who would need to use it, and that adequate assistance is provided for those who may face particular barriers to access.

What does this mean for platforms?
It should be clear on the platform how a user can challenge a decision which has been made to remove content or to suspend their account. Users should always be informed when their content has been removed or their account suspended. When informing the user, clear information should be given on how the user can appeal the decision. Platforms should also consider barriers which may exist for a user to appeal the decision and engage in the grievance mechanism, such as language or disability.

3. Predictable

The principle of predictability requires that there be a clear and known procedure with an indicative time frame for each stage, and clarity on the types of process and outcome available and means of monitoring its implementation.

What does this mean for platforms?
Platforms should set out publicly what the review process is if a user challenges a decision to remove content or to suspend their account. The information should also set out an indicative time frame and what the available remedy (or remedies) will be if the appeal is successful.

4. Equitable

The principle of equity requires that aggrieved parties have reasonable access to sources of information, advice and expertise necessary to engage in a grievance process on fair, informed and respectful terms.

What does this mean for platforms?
Platforms should ensure that users who have had content removed or their account suspended are informed of the full reasons for the decision.

5. Transparent

The principle of transparency requires that parties are informed about the progress of the grievance mechanism and provided with sufficient information about the mechanism's performance to build confidence in its effectiveness.

What does this mean for platforms?
Platforms should ensure that users who appeal against decisions to remove content or suspend their account are informed about the progress of the appeal at regular intervals. It also means that platforms should publish details on the grievance mechanism and how appeals are determined.

6. Rights-compatible

The principle of rights-compatibility requires that outcomes and remedies are consistent with internationally recognised human rights.

What does this mean for platforms?
Platforms should ensure that the available remedies if a user is successful in appealing a decision are effective. Ordinarily, the most effective remedy will be the reinstatement of the content or the account, as the case may be. Depending on the circumstances, other remedies may also be appropriate, such as compensation, a public apology, a guarantee of non-repetition, or a review/reform of a particular policy or process. Remedies should not, themselves, constitute an adverse impact on users' human rights: for example, public apologies about inappropriate or mistaken decisions should not identify the user concerned without their consent, or otherwise interfere with their privacy.

7. A source of continuous learning

The principle of continuous learning requires that that there be regular analysis of the frequency, patterns and causes of grievances to enable the institution administering the mechanism to identify and influence policies, procedures or practices that should be altered to prevent future harm.

What does this mean for platforms?
Platforms should regularly review the frequency, patterns and reasons for appeals against the removal of content or the suspension of accounts, to identify whether any steps need to be taken in reviewing or reforming internal policies and processes to avoid future inappropriate or mistaken decisions.

8. Based on engagement and dialogue

The principles of engagement and dialogue require that there be engagement with affected stakeholder groups about the design and performance of the grievance mechanism, and recommend a focus on dialogue as the means to address and resolve grievances.

What does this mean for platforms?
Platforms should ensure that they engage in regular dialogue with stakeholder groups once the grievance mechanism has been established in order to identify any barriers to continued confidence.

03
Oversight

We believe that there is a middle ground between the current purely self-regulatory approach and the development of national-level regulatory or oversight mechanisms. We propose a new model of oversight which combines industry-developed standards with a multistakeholder mechanism for enforcement.

Overview

- *We propose a new model of oversight based upon the establishment of a global body, the Independent Online Platform Standards Oversight Body (OPSO), funded by online platforms and with a membership determined via a multistakeholder group.*
- *The OPSO would be empowered to assess platforms' compliance with a set of Online Platform Standards (the Standards), developed by platforms following consultation and engagement with relevant stakeholders, and to publish periodic reports with recommendations where appropriate.*

A new model of oversight

The question of whether – and to what extent – a particular sector, industry or profession needs to be regulated is a complex one which requires consideration of many different factors. At one end, there are sectors and services which provide public functions or exercise power or influence such that there is a clear public interest in regulation. Examples include law enforcement agencies or health professionals who may be employed and regulated directly by the government. At the other end, there are sectors and services which are entirely private in nature, or who have a minimal impact upon individuals, meaning that little or no regulation is required, beyond horizontal regulation such as consumer rights or health and safety legislation. Between these two extremes lie a range of different sectors and services which have differing levels of regulation, including self-regulation or co-regulation.

We believe that there is a clear public interest in the activities of online platforms and the services that they provide. As we note in Section 1, many platforms have millions, if not billions, of users, and the services offered are becoming increasingly important and essential in the lives of those users. It is widely accepted that utilities like water, electricity and telephony are recognised as so important to day-to-day life that companies engaged in making them available are not left entirely to market forces and self-regulation.

We believe that there is a clear public interest in the activities of online platforms and the services that they provide.

Increasingly, there is a case for treating the internet – and, by extension, the platforms which make up people's experience of the internet – in the same way. As we also note in Section 1, platforms are becoming increasingly important in enabling individuals to exercise their right to freedom of expression, with the actions of those platforms via content regulation potentially impacting adversely upon that right.

These factors suggest that a purely self-regulatory mechanism is not sufficient to ensure that the interests of users – and the public interest more broadly – is adequately protected. Existing means of accountability for the actions of platforms via investors and stakeholders appear to have little impact. The major voluntary industry-level initiative, the Global Network Initiative (GNI), takes a soft-touch approach – setting out fairly high-level principles in the GNI Principles and Implementation Guidelines, and refraining from publishing full assessments of company members' compliance with them.

As such, we do not believe that the existing mechanisms ensure a sufficient level of protection for the interests of users, including their human rights. While the model we propose in Section 2, if fully implemented, would help ensure a sufficient level of protection for the right to freedom of expression, we judge that pure self-regulation would not provide the necessary transparency, accountability and representation of the public interest. We therefore believe that an additional oversight mechanism should be established to provide that transparency, accountability and representation of the public interest.

We do not, however, believe or propose that such an oversight mechanism should be developed by governments and implemented through national legal or regulatory frameworks. This is for two reasons. First, the global nature of platforms makes national-level mechanisms inappropriate, creating the risk of platforms being forced to comply with scores of different requirements when the issues and interests at stake are global in nature and importance. Second, given the poor human rights record and high levels of censorship in many countries, national level regulation or oversight on issues of content would create significant risks to freedom of expression.

We believe that there is a middle ground between the current purely self-regulatory approach and the development of national-level regulatory or oversight mechanisms. We propose a new, global model of oversight which combines a set of independently developed standards with a multistakeholder mechanism for enforcement. We recognise that there are few, if any, comparable models in other sectors, and that this would be a radical step forward. As such, we have confined our proposals for such a mechanism, at this stage, to relatively high levels of principle, rather than detail.

Developing the oversight mechanism

In the first instance, we propose that interested platforms establish an independent group of experts and set out a Terms of Reference for it to develop the Online Platform Standards (the Standards). The Standards would contain both minimum requirements for platforms as well as an oversight mechanism as detailed below. This group should comprise experts on the relevant issues, including international human rights law, business and human rights, and the operations of platform themselves. In developing the Online Platform Standards, the group of experts should consult with platforms and other interested stakeholders, such as academia, civil society and investors.

We propose a new, global model of oversight which combines a set of independently developed standards with a multistakeholder mechanism for enforcement.

Framework underpinning the new oversight mechanism

We propose that the Terms of Reference should provide for the Online Platform Standards to include the following:

- **Establishment of the OPSO and the Standards**: A global body, the Independent Online Platform Standards Oversight Body (OPSO), would be established, governed by the Standards and by which participating platforms would publicly acknowledge themselves bound. The OPSO would be funded by participating platforms themselves. Any further platform would be able to sign up to the Standards at any time.
- **OPSO membership**: The Standards would set out that membership of the OPSO would comprise a voluntary, multistakeholder group comprising representatives of the platforms, civil society organisations, academia and, potentially, relevant national bodies.
- **Minimum standards**: As well as establishing the OPSO, the Standards would include a commitment from the participating platforms to develop and implement a human rights-respecting framework for content regulation, based on a set of minimum requirements contained within the Standards. These minimum requirements would go beyond the level of principle, and provide detail on the development of Terms of Service, their implementation, and the provision of grievance and remedial mechanisms. We would recommend that our proposed model, set out in Section 2, be considered as the framework, adapted by the platforms as necessary.
- **Standardisation of forms of content**: The Standards could also, where possible, set common categorisations, definitions and understandings of the different forms of unlawful and harmful content which would be subject to restriction. This would promote standardisation and consistency, providing benefits for users themselves when they use multiple platforms, and helping platforms achieve greater efficiency in content moderation and comparison.
- **Support:** The Standards could also provide for platforms to be able to seek advice and assistance from the OPSO on particular issues.
- **Review and amendment**: The Standards would be reviewed periodically (and no less frequently than biennially) to ensure that they remain fit for purpose. Any amendments to the Standards would be developed by independent experts, as with the original Standards, following a process of multistakeholder consultation, including with platforms.
- **Enforcement**: Enforcement of the Standards would be undertaken by the OPSO. The Standards would provide that the OPSO had the authority to assess, at periodic intervals, compliance by the platforms with the Standards. The Standards would require platforms to provide all necessary assistance to the OPSO to be able to carry out its functions, including by providing details on their compliance
- **Transparency**: To improve transparency, the Standards would empower the OPSO to publish reports, and make them publicly available, on compliance by the platforms with the Standards, following each assessment. The reports would also contain recommendations for change to ensure compliance.
- **Non-compliance**: We do not propose that the Standards should give the OPSO any power to sanction platforms for non-compliance with the Standards. Instead, the reports published by the OPSO would contain a clear assessment of whether, and to what extent, the platforms were acting in compliance with the Standards. The reports would also contain recommendations on how non-compliance should be remedied. The Standards could provide for the suspension or expulsion of a platform which repeatedly failed to comply with the Standards.

Although, as noted above, we do not propose any national level regulation of platforms, we nonetheless recognise that there exist a number of national level bodies who have a particular interest in online content regulation. These include national human rights institutions (NHRIs), who have a clear interest in the protection and promotion of human rights at the national level, but also bodies such as the Internet Watch Foundation (UK) and the eSafety Commissioner (Australia) who have mandates to undertake certain functions relating to the regulation of unlawful or harmful content at the national-level. The OPSO should seek to work closely with NHRIs and other bodies with national-level mandates, such as through Memorandums of Understanding.

We also recognise that there are similarities between the model we propose and existing accountability mechanisms, such as the GNI and the Ranking Digital Rights (RDR) Corporate Accountability Index. This model, however, goes further than these existing mechanisms in a number of important respects:

- This model explicitly looks at the issue of content regulation by platforms based on their own Terms of Service whereas the GNI's focus, from a freedom of expression perspective, is on government requests for the removal of content.
- Unlike the RDR Corporate Accountability Index which also includes mobile and telecommunication companies, this model only looks at online platforms, enabling a greater focus on the issues specifically affecting these entities;
- The model gives a voice to certain state actors, such as NHRIs, as well as organisations already involved at the national level in the issue of content regulation by online platforms.
- The model would include the publication of the compliance reports prepared by OPSO, providing greater transparency on the practices of online platforms and their shortcomings, whereas the full assessments and evaluations of compliance by online platforms with the GNI principles are not currently published.

04
The role and responsibilities of state actors

As we note at the start of this white paper, our focus is on the development of a rights-respecting model of online content regulation by platforms. We have deliberately not proposed any legislative or regulatory model of online content regulation established or administered by state actors.

Although we propose that certain national bodies, such as NHRIs, be involved in the oversight mechanism, the OPSO, that is the limit of our proposed involvement of state actors in the model. However, platforms do not operate in a vacuum. The national legislative and regulatory frameworks in the states in which they operate place limitations – or establish obligations – with which they must comply. Such limitations or obligations, were they to impact upon the ability of platforms to establish the model we propose, would risk undermining the effectiveness of the model. As such, it is essential that states ensure that their legislative and regulatory frameworks, as they apply to platforms, do not hinder their ability to establish the model we propose, or undermine it in practice. We do, however, see a role for governments in taking steps to encourage platforms to increase their adherence to the principles which underlie the model.

Such a position is entirely consistent with the negative obligation of states not to restrict the right to freedom of expression (as well as other human rights, of course) including as it is exercised online, save where such restrictions are required or permitted under international human rights law. There are two particular ways that legislative and regulatory frameworks could breach this obligation as well as interfere with the ability of platforms to establish the model we have proposed in Section 2.

The first is by prohibiting, through legislation or otherwise, certain forms of content where such prohibitions apply to online platforms but which cannot be justified under international human rights law. The second is by imposing inappropriate liability on platforms for the content which they make available.

It is essential that states ensure that their legislative and regulatory frameworks, as they apply to platforms, do not hinder their ability to establish the model we propose, or undermine it in practice.

General restrictions on content

States have a general duty not to restrict the right to freedom of expression save in a number of very limited circumstances. Some of these circumstances are mandated by particular human rights instruments, such as Article 20 of the International Covenant on Civil and Political Rights (ICCPR) (propaganda for war and any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence) and the Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography (images of child sexual abuse). The other circumstances in which states may restrict freedom of expression are set out in Article 19(3) of the ICCPR, namely where it is "provided by law" and necessary for respect of the rights or reputations of others, or for the protection of national security or of public order, or of public health or morals.

States should therefore ensure that any general restrictions which apply to online content are consistent with their obligations with respect to the right to freedom of expression, and repeal or amend legislation, or review practices, which are inconsistent.

Restrictions specific to platforms

In addition to ensuring that the broader legal framework, where it touches upon freedom of expression as it is exercised online, is consistent with international human rights law and standards, states should ensure that any legislation which is of specific application to platforms does not restrict freedom of expression explicitly or in its effects. Of particular concern is legislation which attaches liability to platforms for content which is available on them, which can lead to a 'chilling effect' in which platforms either become reluctant to host or otherwise make available content, or are overly zealous in removing content which might be harmful.

There are, at present, a range of liability regimes which fall within three broad categories, outlined in the below table.

States have a general duty not to restrict the right to freedom of expression save in a number of very limited circumstances.

Liability regime	Summary	Examples
Strict liability	Platforms are held liable for unlawful or harmful content made available by users on their platforms, even if they are not aware of the content.	Thailand (Section 15 of the Computer Crimes Act 2007)
Conditional liability/ 'safe harbour'	Platforms are not held liable for unlawful or harmful content made available by users on their platforms provided they do not have any knowledge of the content or, if they do have knowledge, have acted expeditiously to remove that content.	European Union (Article 14 of the E-Commerce Directive)
Broad immunity	Platforms are, as a general rule, not held liable for unlawful or harmful content made available on their platforms, even if they are aware of the content. Some limited exceptions may exist, such as for certain specified crimes or intellectual property.	USA (Section 230 of the Communications Decency Act)

'Strict liability' regimes are the most likely to result in overly broad restrictions of freedom of expression, as they require the platform proactively to monitor and remove content, even without notification. However, even 'safe harbour' or 'conditional liability' regimes can be problematic particularly where the conditions under which liability will be held are such that they require a platform to make determinations about the lawfulness of content, to remove content within short time limits or impose high sanctions for a failure to take down content. In such circumstances, there is a clear incentive on platforms to 'play it safe' and remove ambiguous content so as to avoid liability and potential fines or other sanctions. One example of such a liability regime is the recently adopted Network Enforcement Act (NetzDG) in Germany. The NetzDG requires platforms with more than two million subscribers to remove "manifestly unlawful" content within 24 hours with fines of up to €50 million for non-compliance.

While we do not consider that intermediaries should never be liable for content which is made available on their platforms, we consider that there must be sufficient limitations and safeguards in place when it comes to attaching liability to ensure that risks to freedom of expression through incentives to remove content are effectively mitigated. We believe that such a regime is feasible through compliance with the following principles:

- First, the development of any legislation which attaches liability to platforms should be open, inclusive and transparent. The development process should include consultation with all relevant stakeholders and states should consider undertaking a human rights impact assessment to understand the impact that the legislation may have on human rights.
- Second, the legislation itself should be consistent with the principle of legal certainty. This means that it should be accessible, and sufficiently clear and precise for platforms, users and other interested groups to be able to regulate their conduct in accordance with the law.
- Third, the legislation should not directly or indirectly impose a general obligation on platforms to monitor third party content where they do nothing more than host that content, or transmit or store it, whether by automated means or not. Further, the legislation should not attach strict liability to a platform for hosting unlawful content as this would, de facto, require such monitoring.
- Fourth, the legislation should not directly or indirectly impose liability on platforms for third party content where they do nothing more than host that content, or transmit or store it, whether by automated means or not, and have no actual knowledge of specific content thereby hosted, transmitted or stored. Indeed, the legislation should explicitly exempt platforms from liability in such circumstances.
- Fifth, the legislation should not attach liability to platforms for failing to restrict lawful content.
- Sixth, the legislation should not provide any incentives to remove content which may be lawful, such as via unrealistic timeframes for compliance, or the imposition of disproportionate sanctions for non-compliance.

While we do not consider that intermediaries should never be liable for content which is made available on their platforms, we consider that there must be sufficient limitations and safeguards in place when it comes to attaching liability to ensure that risks to freedom of expression through incentives to remove content are effectively mitigated.

References

1. There are, of course, many other models and proposals for how platforms should respond to the issues raised in this white paper. See, for example, Newland, E., Nolan, C., Wong, C., and York, J., "Account Deactivation and Content Removal: Guiding Principles and Practices for Companies and Users", *The Berkman Center for Internet & Society and The Center for Democracy & Technology,* September 2011; Article 19, "Self-regulation and 'hate speech' on social media platforms", 2018; the Ranking Digital Rights Corporate Accountability Index; and the UN-mandated Tech Against Terrorism initiative.

2. International Telecommunication Union, *ICT Facts and Figures 2017,* 2018, p. 2, available at: https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf/.

3. *Ibid.*

4. Titcomb, J., "Facebook now has 2 billion users, Mark Zuckerberg announces", *The Telegraph,* 27 June 2017.

5. Bergman, S., "We Spend A Billion Hours A Day on YouTube, More Than Netflix And Facebook Video Combined", *Forbes,* 28 February 2017.

6. Twitter, "How Policy Changes Work", *twitter.com,* 20 October 2017, available at: https://blog.twitter.com/official/en_us/topics/company/2017/HowPolicyChangesWork.html.

7. United Nations Human Rights Council, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression,* UN Doc. A/HRC/32/38, 11 May 2016, Para 2.

8. Article 14 of the Directive on electronic commerce (Directive 2000/31/EC), for example, provides that service providers should not be held liable for content hosted unless (a) they have "actual knowledge" of its illegal nature or (b) upon obtaining such actual knowledge, they fail to act expeditiously to remove or to disable access to the content.

9. The Global Network Initiative is a multi-stakeholder group of companies, civil society organizations, investors and academics which developed the Global Principles on Freedom of Expression and Privacy and accompanying Implementation Guidelines, with the aim of protecting and advancing freedom of expression and privacy in the ICT sector. More information can be found at www.globalnetworkinitiative.org.

10. Venturini, J. and others, "Terms of Service and Human Rights: an Analysis of Online Platform Contracts", Editora Revan, 2016.

11. *Ibid.,* p. 96.

12. Ranking Digital Rights, *Corporate Accountability Index 2017: Key Findings,* available at: https://rankingdigitalrights.org/index2017/findings/keyfindings.

13. Browne, M., "YouTube Removes Videos Showing Atrocities in Syria", *The New York Times,* 22 August 2017.

14. BBC News, "Qatar's Al Jazeera Twitter account back after suspension", *www.bbc.co.uk,* 17 June 2017, available at: http://www.bbc.co.uk/news/world-middle-east-40311882; Solon, O., "Two cases of Twitter abuse highlight the obscure nature of suspensions", *The Guardian,* 10 January 2017.

15. TIME, "The Story Behind the 'Napalm Girl' Photo Censored by Facebook", *time.com,* 9 September 2016, available at: http://time.com/4485344/napalm-girl-war-photo-facebook.

16. Nielsen, N., "Commission: 120 minutes to remove illegal online content", *euobserver.com,* 9 January 2018, available at: https://euobserver.com/justice/140482.

17. BBC News, "Call for tech giants to face taxes over extremist content", *www.bbc.co.uk,* 31 December 2017, available at: http://www.bbc.co.uk/news/uk-42526271.

18. Stewart, H., "May calls on internet firms to remove extremist content within two hours", *The Guardian,* 20 September 2017.

19. Chrisafis, A., "Emmanuel Macron promises ban on fake news during elections", *The Guardian,* 3 January 2018.

20. Article 19 is, itself, inspired by Article 19 of the Universal Declaration of Human Rights, adopted in 1948.

21. UN Human Rights Committee, *General Comment No. 34: Article 19: Freedoms of opinion and expression*, UN Doc. CCPR/C/GC/34, 12 September 2011, 2011, Para 11.

22. *Ibid.,* Para 12.

23. Data taken from Google, Government requests to remove content, available at: https://transparencyreport.google.com/government-removals/overview.

24. Data taken from Google, YouTube Community Guidelines enforcement, available at: https://transparencyreport.google.com/youtube-policy/overview.

25. See above, note 21.

26. See, for example, United Nations, Human Rights Council, *Resolution 32/12. The promotion, protection and enjoyment of human rights on the Internet,* UN Doc. A/HRC/RES/32/13, 18 July 2016.

27. See above, note 21, Para 25.

28. UN General Assembly, *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN Doc. A/66/290, 10 August 2011. See, in particular, Paras 20-36.

29. See above, note 24

30. Center for Democracy & Technology, *"Mixed Messages? The Limits of Automated Social Media Content Analysis",* 28 November 2017, available at: https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis.

31. See, for example, Article 19, "EU fails to protect free speech online, again", article19.org, 5 October 2017, available at: https://www.article19.org/resources/eu-fails-to-protect-free-speech-online-again/.

32. See, for example, Article 2(3)(a) of the ICCPR which requires states to ensure that any person whose rights of freedoms are violate has an effective remedy.

GLOBAL PARTNERS DIGITAL

Second Home
68 Hanbury St
London E1 5JL

+44 203 818 3258
info@gp-digital.org