Content moderation in user-to-user online services

An overview of processes and challenges

Report

Published 14 September 2023

Contents

Section

1. Overview

Introduction

1.1 **Concerns about harms related to social media and other online services that host user-generated content (user-to-user services) have become a focus of public debate in recent years, both in the UK and globally.** Many of these concerns fall under Ofcom's remit through our media literacy duties and powers under the video-sharing platform regime, or are likely to fall within our duties under the Online Safety Bill.

1.2 **Online service providers engage in a wide range of activities aimed at limiting users' harm – commonly referred to as Trust and Safety (T&S) activities.** These include activities aimed at preventing bad or unwanted *actors* (e.g. criminals) from abusing services; at tackling undesirable *behaviour* (e.g. trolling); at tackling illegal or harmful *content* (e.g. hate speech); and at preventing children from *accessing* certain content or services.

1.3 **An important part of content-related T&S activities revolves around removing or reducing the visibility of potentially harmful content. In this paper we refer to these activities as *content moderation*.** Content moderation is central to current public discussions about online regulation, in part because it raises implications for how users can express themselves freely online. It will also be relevant to Ofcom's future work on online services' safety systems and processes. However, information in the public domain on content moderation is relatively scattered and incomplete, and it may be difficult for non-experts to form a holistic view of common practices and challenges. This paper aims to help address this situation, drawing on extensive discussions we held over the last two years with six service providers of different sizes and types, including Facebook (Meta), YouTube (Google), Reddit and Bumble – as well as information in the public domain.

1.4 **This report consists of two Parts.** Part I provides a factual description of the main aspects of content moderation, relying on a simplified and stylised account that we believe is broadly reflective of services' approach to content moderation – even if it may not exactly reflect the situation at any one service. Part II presents our own reflections on some of the challenges involved in content moderation, as well as on recent efforts to develop metrics to track services' performance in content moderation.

Part I: The content moderation process

1.5 **Content moderation relies on general rules, or content *policies*, that in principle apply to *all* content. Policies are applied to individual content items at scale through *enforcement processes*.** Below we discuss, first, content policies and the work involved in developing these (policy-setting); and second, enforcement processes.

Policy-setting

1.6 **Policies specifying what types of content are not allowed on a service are fundamental to service providers' safety efforts.** In this document we refer to these as *content standards*, and to content prohibited by them as *violative* content. Generally, content standards aim to prohibit all illegal content plus any other content that service providers consider harmful to their users or otherwise undesirable on their service.

1.7 **It is important to note that service providers normally aim to remove violative content *when they become aware of its nature*,** for example after a human moderator has reviewed the content.

1.8 **In addition to banning violative content, some services target certain non-violative problematic content with non-removal actions, such as reducing the visibility of content.** Much of the targeted content is currently at the centre of public debate and – depending on the service – may include:

a) 'borderline-violative' content that comes close to breaching services' standards without quite doing so;
b) potentially violative content that has been reported by a user or flagged by an automated tool as such, but which has not yet been reviewed by a human moderator; and
c) other types of problematic content, such as sensationalistic content, conspiracy theories, or clickbait.

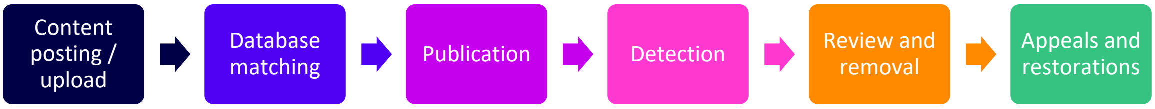
1.9 **In recent years some service providers have begun to publicly disclose key aspects of their treatment of non-violative problematic content.** However, publicly available information on this is generally less detailed than information relating to the content standards that lead to content removals.

Policy enforcement

1.10 **Once content standards are set, service providers rely on technical systems, personnel and processes to enforce policies at scale.** The preparatory steps for this can include developing detailed moderation guidelines to help moderators, setting out how policies should be applied in a variety of situations; training of human moderators; and training of automated 'classifiers'. With this preparation complete, service providers apply their policies to content uploaded and published on their services.

1.11 **A stylised process for removing violative items can be conceptualised in terms of the life-cycle of a content item posted online**, as summarised in Figure 1:

Figure 1: A stylised content moderation process



1.12 The key steps of this process are the following:

a) **Content posting / upload**: a user submits a content item (e.g. video or a piece of text).
b) **Database matching**: automated systems compare the uploaded content against databases of known violative content and, if a match is found, prevent publication.
c) **Publication:** content that passes the matching stage is made available to other users, typically seconds after posting.
d) **Detection of potentially violative content**: this may arise from

- **Flagging by AI-based 'classifiers'** identifying potentially violative items, or
- **Reports by end-users** or third-party organisations who have seen the content.

e) **Review and removal**: Content flagged by classifiers and/or reported by users as being potentially violative is sent to human moderators for review (unless content is removed automatically, in which case human review may or may not take place). Of note,

   - There is typically a time-lag between content being referred and it being reviewed by moderators, due to resource constraints and the potentially large and fluctuating volume of potentially violative items referred.
   - Because of this, services commonly use algorithms to ensure that reviewers' in-trays are prioritised in such a way as to minimise overall harm in some sense (e.g. by prioritising content that is likeliest to be most harmful and/or to be viewed by the largest number of people).
   - If human reviewers find that an item violates a service's standards, it will be removed (or 'taken down'). Users responsible for posting violative content may be penalised, for example through a temporary ban on future posting.

f) **Appeals and restorations:** Typically users whose content is removed can submit an appeal. Decisions not to remove an item may also be appealable.

1.13 **The stylised process described above is not universally applicable.** Of note:

a) The sequencing of automatic detection and publication may vary: automated detection of potentially violative content may occur before or after publication.

b) In some cases, detections of potentially violative content may be followed by automatic removals (this is common for spam, but may also be applied more broadly).

c) Some services rely on community-moderation approaches involving volunteers who may be granted a degree of responsibility for enforcing standards.

d) Services providing private (one-to-one) communications not widely visible to other users may rely on different approaches to reducing harm, such as automatic filtering or enforcement action targeted at non-compliant users.

e) Some service providers choose to have human moderators review all content prior to publication (a practice used in some high-risk contexts); however, this is relatively rare.

1.14 **In addition to content removals, as noted earlier some service providers take measures to reduce the visibility of various types of potentially problematic content that is not necessarily violative**. Measures include:

a) **Down-ranking (or 'demoting')** content, so that it appears less frequently in services' from pages, users' news feeds and lists of recommended content. Depending on the service and context, demotions may be applied in a personalised way, so that an item may be demoted for some users but not for others.

b) **Overlays (or 'interstitials' or 'panels'), content blurring, and accompanying 'labels'** – aiming to ensure that users who access certain problematic or controversial content are aware of essential context, such as the disputed nature of claims, and/or are offered links to authoritative sources of information.

c) **Other interventions**, including prompts encouraging users to pause and think before posting content that appears potentially problematic; age-gating to restrict access to certain content or versions of a service; and restrictions on posting, sharing and forwarding functions, for example if a post has been shared a large number of times.

Part II: Our reflections

Challenges and trade-offs in content moderation

1.15 **As the preceding discussion illustrates, in general content moderation processes are designed to limit the viewing of violative content, rather than to guarantee that no harmful or violative content can be accessed.** For example, under widely adopted content moderation practices, human moderators only see content that has been previously identified by AI classifiers or reported by end users as being potentially violative, and violative content may remain available until it is reviewed by a human moderator.

1.16 **These limitations reflect decisions in process design, and these decisions may entail trade-offs which may or may not be explicit**. For example, a decision to allow items reported as potentially violative to remain 'live' until a moderator sees them may lead to fewer non-violative items being removed erroneously than if such items are automatically removed pending review – but potentially with the result that violative items remain available for longer, and are viewed by more users, than otherwise. Thus, this decision may be seen as entailing a trade-off between enabling users to express themselves freely on the service, and keeping users safe.

1.17 **Within the constraints set by the overall structure of their processes, service providers can and do rely on a variety of levers to reduce harm.** In the example above, the service may be able to reduce the 'turnaround time' between content upload and review (and removal) by hiring more moderators, thereby reducing the amount of time that potentially violative content is 'live'; but the more moderators are employed, the higher the costs of running a service. More broadly, although service providers have a range of levers at their disposal to reduce harm, these generally entail costs and trade-offs, and none offers a silver bullet. We set out some of these levers and trade-offs in Box 1.

Box 1: Selected decisions and trade-offs

How broadly or narrowly to define violative content

In some cases, services may prohibit broad content categories capturing both harmful and potentially harmless content. However, broad prohibitions may be controversial on services or in contexts where users may expect a degree of unrestricted expression, or where the content could have, for example, cultural or educational value. Alternatively, policies more narrowly targeting harmful types of content may risk allowing more harm, for example through 'borderline' content that is not removed.

How aggressively to apply demotions and other non-removal measures

Some items may show signals of being potentially problematic while *also* showing signals of being potentially engaging and/or of being valued by certain users. While the first type of signal might normally lead a service to demote an item, the second type might normally lead to items being *promoted*, at least for some users. How service providers calibrate their systems determines whether such items are ultimately promoted or demoted for a given user. In turn this may impact not only on the degree to which harm is reduced *but also* on the degree to which users are prevented from seeing content which may particularly interest them and which may not be violative.

How to prioritise items awaiting moderation

Systems can be designed to prioritise moderators' review of potentially violative items according to different criteria, such as the popularity / virality of an item, the seriousness of the suspected harm, or the likelihood that the item will be confirmed as violative. Depending on which criteria take precedence, harm may be reduced in different ways – e.g. many users might be prevented from seeing not-particularly-harmful content, or alternatively a smaller number of users might be prevented from seeing particularly harmful content. Trade-offs of this type may be unavoidable in a context of finite human moderation capacity.

How much human moderation capacity to have

The trade-off around prioritisation above might potentially be made less acute by making moderation faster, so that *all* items, whether they are particularly harmful or are 'going viral', can be reviewed more quickly. One way to do this may be to employ more human moderators. However, this comes at a cost, which may be substantial. Moreover, services may consider that the return on investment (in the sense of harm reduction) from continued expansion of moderator capacity may diminish as extra moderators tackle comparatively less harmful content, as a result of effective automatic prioritisation.

To what extent to rely on AI-based classifiers to make decisions

Automatic classifiers can play a key role in content moderation, not only helping prioritise moderators' work but also demoting and, in some cases, removing content automatically. However, automatic moderation systems can make mistakes. Service providers currently must decide what rates of 'false negatives' and 'false positives' to accept in relation to demotions and automatic removals, as well as whether an increase in one of these two types of error is an acceptable trade-off for a reduction in the other.

Whether and when to implement pre-moderation/screening

While human moderation of all content prior to publication is rare in the industry, it is another design decision services may make, which might reduce exposure to violative content significantly, but potentially at a substantial financial cost, while potentially being perceived negatively by users.

Towards performance metrics for content moderation

1.18 **Inasmuch as content moderation can only reduce but not eliminate all harm, its success is a matter of the *degree* and the *sense* in which harm is reduced.** A key question is therefore how success might be measured. Some service providers regularly publish a range of quantitative metrics as part of their periodic transparency reports, including the number of items removed, how long violative items stay online before being removed, and how many removals were appealed.

1.19 **In recent years some service providers have introduced metrics reflecting the *viewing* of violative content, which some providers see as particularly important –** describing them as "the number we hold ourselves accountable to"[1] or "the primary metric [we use] to measure our responsibility work".[2] This includes Facebook and Instagram's 'prevalence' metric and YouTube and Snap's 'violative view rate'.

1.20 **We welcome service providers' contributions to a richer public understanding** not only through the availability of these metrics, but also, through interventions like those quoted above, which draw attention to the question of what measurable outcomes content moderation should be measured by.

1.21 **While the available data is certainly valuable, certain voices in the expert community have called for service providers to publish more granular information**, for example relating to the concentration of exposure to violative content among certain groups of users, or to exposure to borderline content or content targeted by visibility-reducing measures. In our view, greater transparency can empower and inform users, researchers, investors and other parties, and public understanding would benefit from further information on how services measure the success of content moderation.

[1] See the explainer published by Facebook (2018), 'Understanding the Facebook Community Standards Enforcement Report.

[2] See YouTube's blog post (2021), 'Building greater transparency and accountability with the Violative View Rate'.

2. Introduction

2.1 **Concerns about harms related to social media and other online services that host user-generated content (user-to-user services) have become a focus of public debate in recent years, both in the UK and globally.** These concerns relate, among other things, to harmful content disseminated through online services, particularly content that is illegal or harmful to children; to harmful interactions between users; to the potential market power of certain firms; and to the handling of users' personal data. Many of these concerns fall within Ofcom's remit. We have duties to promote media literacy in order to further the interests of citizens and consumers; since November 2020 we have been the regulator of UK-established video-sharing platforms (VSPs) in relation to online content; and the Online Safety Bill will give Ofcom a wide set of duties relating to the protection of users from online harm (Box 2).

Box 2: the Online Safety Bill

The Online Safety Bill proposes that Ofcom will oversee a new independent regulatory regime to ensure that online services assess and mitigate the risks to their users' safety, especially relating to illegal content and the protection of children. The planned rules won't involve us regulating or moderating individual pieces of content. Ofcom will have powers that include:

- overseeing the requirements on services to assess risks on their services and to have proportionate systems and processes (including content moderation processes) to protect users and individuals from illegal content;
- overseeing additional requirements on relevant services to assess risks to children and to put in place proportionate protections for children against harmful content (which again may include content moderation);
- holding 'Category 1 services' (the user-to-user services who meet the relevant thresholds) to account for additional duties including the consistent application of their own terms and conditions; and
- robust information gathering and enforcement powers, alongside transparency reporting powers for some services, which will help Ofcom inform the public about how those services are protecting their users.

To support services in complying with these duties Ofcom is required to produce a range of Codes of Practice and guidance for regulated services.

2.2 **Online service providers engage in a range of activities aimed at limiting users' online harm – commonly referred as Trust and Safety (T&S) activities.** These include[3] activities aimed at preventing bad or unwanted *actors* (e.g. criminals) from abusing services; at tackling undesirable *behaviour* (e.g. trolling); at tackling illegal or harmful *content* (e.g. hate speech); and at preventing children from *accessing* certain content or services. Service providers' actions as part of T&S include closing or suspending accounts that violate services' rules; making potentially harmful content unavailable or less visible; limiting access to

[3] We borrow the categorisation of services' activities into those centred on actors, behaviour or content (ABC) from *Freedom and Accountability: a transatlantic framework for moderating speech online*, the final report of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (2020).

certain content to some users (e.g. adults only); limiting some users' ability to post content; introducing or highlighting helpful content (e.g. links to health authorities); and relying on 'safety by design' (whereby functionality is designed to encourage or discourage certain behaviours, such as e.g. impulsive 'reposting' of content after only reading a headline; or to make certain undesired experiences opt-in, such as e.g. seeing explicit messages).

2.3 **These activities can work together to tackle certain harms**; for example, disinformation posted by a known malicious actor might be removed on account of it falling foul of content rules, or alternatively on account of its author displaying 'inauthentic coordinated behaviour', while its viral spread may be slowed down by 'friction' measures designed to discourage the sharing of inflammatory content.

2.4 **In this report we focus primarily on services' content moderation activities – by which we mean all efforts to prevent, remove and reduce the visibility of certain content.** Content moderation is central to current public discussions about online regulation, in part because it raises implications for how users can express themselves freely online. It will also be relevant to Ofcom's future work on services' online safety systems and processes. However, information in the public domain on content moderation is relatively scattered and incomplete, and it may be difficult for non-experts to form a holistic view of common practices and challenges.[4]

2.5 **To help develop our understanding of content moderation, over the last two years we worked with six service providers of different sizes and types, including Facebook (Meta), YouTube (Google), Reddit and Bumble**, focusing particularly on how providers can identify, tackle and track harm.[5] Service providers engaged on a strictly voluntary basis and we are grateful for the time and effort they devoted to this work. This report presents a summary of our findings from these experiences.

2.6 **This report consists of two Parts:**

a) **Part I provides a factual description of the main aspects of content moderation.** We rely on a simplified and stylised account that we believe is broadly reflective of services' approach to content moderation, even if it may not exactly reflect the situation at any one service. We focus primarily on social media services and similar contexts where users can post content that can be viewed by large numbers of other people, and less on cases where users send messages to a single recipient or a small group of recipients. We discuss policy design in section 3 and policy enforcement in section 4.

b) **Part II presents our own reflections on some of the challenges involved in content moderation, as well as on recent efforts to develop metrics to track services' performance in content moderation.** In section 5 we note that, as currently

[4] Industry- or practitioner-led efforts to bring together Trust and Safety expertise include the Integrity Institute, the Trust & Safety Professional Association, the https://dtspartnership.org/, and the Trust & Safety Teaching Consortium. For academic perspectives see Keller, D. and Leerssen, P., 2020. 'Facts and where to find them: Empirical research on internet services and content moderation'. *Social media and democracy: The state of the field and prospects for reform*, p.220-251; Goldman, E., 2021. 'Content Moderation Remedies', *Michigan Technology Law Review*, 28, p.1-59; Gerrard, Y., 2022. 'Social Media Moderation: The Best-Kept Secret in Tech' in *The Social Media Debate*, p. 77-95. Routledge; and Douek, E., 2022. 'Content Moderation as Systems Thinking', *Harvard Law Review*, 136, p.526-607. Ofcom has also previously published a general model of the structures and workflows of online user-to-user services (the 'A-SPARC model', 2021), which includes aspects of content journeys and content moderation.

[5] Our central focus was on understanding the impact of content moderation on key measurable outcomes that services can track.

implemented, in general content moderation can reduce but not eliminate harm, and that while service providers use a number of levers to reduce harm, each of these comes with costs and/or trade-offs. In section 6 we reflect on recent efforts by service providers and other stakeholders to identify metrics to track the performance of content moderation.

2.7 **This report is based on discussions with the service providers mentioned above, as well as information in the public domain** published by these as well as by a wider range of other service providers.[6] Non-public information is used with providers' consent and without attribution, to protect confidentiality.

2.8 **Although this report is largely based on our engagement with service providers, Ofcom is its sole author.** While we have asked the service providers named above to comment on the accuracy of Part I, we have not asked them to approve or endorse this text. Service providers were not given an opportunity to comment on our reflections in Part II. Services' approaches to content moderation are constantly evolving and may have changed since our conversations with them and/or since they published the documents that we reference in this report.

[6] For the avoidance of doubt, our footnotes reference relevant information in the public domain from a range of service providers, including from many that did not participate in the work behind this report.

Part I: Description of the content moderation process

3. Policy-setting

3.1 **Content moderation revolves around general rules, or *policies*, that apply to *all* content** on a service. Thus, for example, a service might prohibit content of a 'graphic violent' nature, with a specific meaning set out in its policy. Most key policies are public; that is, they are set out in writing in web pages accessible to users, and users are expected to abide by these. All policies are *forward-looking* in the sense that they apply to content that may not yet exist. Policies may reflect a service provider's wider publicly-articulated mission or values,[7] which may go beyond issues of trust and safety (e.g. services may place value on authenticity, free expression, etc).

3.2 **All service providers we worked with have policies aimed at preventing user harm, including provisions designed to ensure that illegal content is explicitly banned**. Approaches to developing and evolving harm-reduction policies vary across services and include responding to insights that may emerge from enforcement processes or from specialist teams and contractors tasked with detecting new patterns of problematic content or behaviour; responding to external events, such as the rise of misinformation and disinformation during the Covid-19 pandemic; and collaborating with industry bodies[8] and external experts.

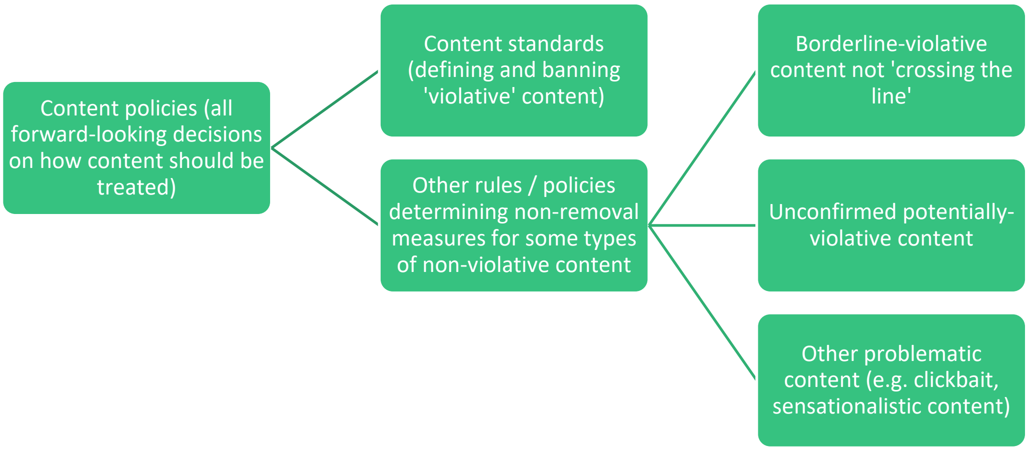
3.3 **Content policies can be categorised under two broad groups**: (i) *content standards,* which ban certain types of content, referred to as 'violative' content; and (ii) other policies that target certain problematic but non-violative content with measures that fall short of banning or removal, such as making content harder to find. While content standards and removals are often publicly documented in detail, the treatment of non-violative content through non-removal measures is generally less well documented.[9] We summarise these concepts in Figure 2 for ease of exposition, noting that our terminology may not always be in line with the relevant literature. In this section we focus only on how these different types of content can be *defined* for policy-setting purposes; we discuss the *treatment* of different types of problematic content ('enforcement') separately, in section 4.

[7] Examples include: within Meta's Community Standards; About YouTube; Reddit Community Values; X Security and Privacy; X Civic Integrity; OnlyFans Mission, Vision and Values; within Nextdoor's Community Guidelines; and within Xbox's Community Standards.

[8] For instance, the Global Internet Forum to Counter Terrorism (GIFCT) is a multistakeholder group specific to this particular issue.

[9] Moreover, when services do publish information on non-removal measures, these may not be described as 'policies' at all. However, just like for content standards, in introducing non-removal measures services make forward-looking decisions applying to wide ranges of content, driven at least partly by the need to protect users. For these reasons, for ease of exposition here we refer to these measures as implementing a certain kind of 'policy'.

Figure 2: A taxonomy of content policies


Content policies (all forward-looking decisions on how content should be treated)
Content standards (defining and banning 'violative' content)
Other rules / policies determining non-removal measures for some types of non-violative content
Borderline-violative content not 'crossing the line'
Unconfirmed potentially-violative content
Other problematic content (e.g. clickbait, sensationalistic content)

Content standards and ‘violative’ content

3.4 **Policies specifying what types of content are not allowed on a service are fundamental to service providers’ safety efforts.** We refer to these policies as ‘content standards’. These normally form part of a service’s publicly-facing terms of service and have names such as ‘community standards’ (Facebook), ‘community guidelines’ (YouTube and Bumble) or ‘content policies’ (Reddit). Items that breach these standards are commonly referred to as ‘**violative**’ content. Users posting content are expected to understand and observe these rules.

3.5 **Generally, content standards aim to prohibit all illegal content, plus any other content that service providers consider harmful to their users or otherwise undesirable**. Common examples include graphic violence, incitement to violence and hate, harassment, self-injury, nudity, and spam. Not all violative content or behaviour need be illegal or even harmful; service providers may deem certain types of content as simply inconsistent with their wider values or disruptive of users’ experience of the service (e.g. spam may be annoying but not necessarily harmful).

3.6 **Conversely, it is not guaranteed that all content that individual users or other stakeholders might consider harmful will be prohibited**. What ‘harmful’ means, and what content may be considered ‘harmful’, is open to multiple interpretations, partly because users themselves may have different perceptions of harm and tolerance for potentially harmful content.

3.7 **It is important to note that, in prohibiting certain types of content, service providers usually aim to remove this content *after* they have become aware of it and have determined its nature**, for example following a report by users or other third parties. However, service providers may not necessarily undertake to prevent all instances of violative content from ever being available on their services. We discuss policy enforcement in section 4.

Other policies

3.8 **Some types of controversial content currently at the centre of public debate are often *not* violative**. For example, misinformation and conspiracy theories are often not *per se* banned categories under services’ content standards. However, if individual instances of such content fall foul of existing content policies (e.g., policies concerning racial hatred), such content would be violative and would be removed if found.

3.9 **Some service providers target certain types of problematic non-violative content with non-removal measures aimed at limiting the visibility of content, or at informing users that the content may be problematic.** Depending on the service, this may include:[10]

a) ‘**borderline**-**violative’** content that comes close to breaching services’ standards, without quite doing so;[11]

b) **potentially violative** content that has been identified (by AI-based classifiers or by service users) as being potentially violative but which has not yet been confirmed as such through the review process (described below); and

c) **other types of problematic content**, such as sensationalistic content, conspiracy theories, or clickbait, which may not necessarily be suspected of violating a service’s standards, but which service providers (or users) may deem undesirable, or of low value.

The measures that may be deployed in relation to this content include lowering the visibility of items (‘demotion'), introducing interstitial panels, and introducing ‘friction’ on onward-sharing functionality. We discuss non-removal measures starting at p 22.

3.10 **Until recently, only limited information existed in the public domain on what content may be targeted by non-removal measures, or on the range of non-removal measures that service providers use**. Although some service providers have started publishing information about non-removal measures,[12] to date this is generally less detailed than information relating to the content standards that lead to content removals.

[10] These categories do not necessarily map to categories and terminology used by the services. For a description of types of content targeted by YouTube, see ‘On YouTube’s recommendation system’ (2021). For Facebook’s equivalent see ‘Types of content that we demote’ (updated 2023). Other services publish similar explanations such as X, ‘Our range of enforcement options and ‘How TikTok recommends videos #ForYou’ (2020).

[11] Facebook has published a description of several borderline content types.

[12] See, for example, footnote 10 above, as well as information published by Instagram to help users understand why their content may sometimes not be included in recommendations presented to other users.

4. Policy enforcement

Policy rollout

4.1 Content policies may be applied to millions of content items every day. Service providers may therefore need to prepare their systems, personnel and processes to work at scale. Depending on the service, this may include:

a) **Development of detailed moderation guidelines:** While the policies that service providers publish aim to give users a clear sense of what content is banned, they may lack enough detail to ensure that less clear-cut cases are treated consistently. Accordingly, some service providers produce more detailed, internal moderation guidelines for their moderators, with more definitions, exceptions and examples (and, in some cases, examples specific to a certain country's laws or cultural context). These guidelines may remain unpublished.[13]

b) **Training of human moderators:** Staff (who may be employed by a service provider or by a subcontractor) are trained to apply policies consistently and objectively.[14] In certain cases, specialist moderator teams are tasked with dealing with specific policies (harms). The training, management and well-being of moderation teams are important issues that services need to consider while implementing and enforcing their content policies; while they are not the focus of this report they have been considered by academic, media and industry commentators, especially in relation to the wellbeing risks for content moderators and the potential mitigations to those risks.[15]

c) **Training of automated systems:** As we discuss below, automated systems known as 'classifiers' often complement human moderators by identifying suspect items and in some cases removing violative items automatically. Classifiers must be 'trained' with examples of content that have been classified previously by humans, so that they can 'learn' to make similar decisions when presented with new content.[16] Services also use technology to detect copies of known violative content.

4.2 Depending on the service, some or all of the above may be done when new policies are introduced and/or when new moderator personnel or technical systems are introduced. As these activities progress, more and more violative items will be caught, and fewer false positives will be wrongly removed or unnecessarily sent for review.

[13] Ofcom recently found that, among regulated Video-Sharing Platforms (none of which was involved in the discussions behind the present report), moderators do not always have sufficient guidance on how to enforce VSPs' terms and conditions. See our report, 'Regulating Video-Sharing Platforms (VSPs). Our first 2023 report: What we've learnt about VSPS' user policies' (2023).

[14] The training of moderators is mentioned in the Video-Sharing Platforms report in the previous footnote. Meta has also published some brief information about how they train and give feedback to their teams.

[15] For an academic discussion of the issues, see e.g. Steiger, M. et al. 2021. 'The psychological well-being of content moderators', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, p.1-14; for media coverage, see e.g. BBC, 'Facebook moderator: "every day was a nightmare"'(2021); for a service provider's perspective see e.g. Google Wellness Standards for Sensitive Content Moderation.

[16] For more on classifiers and their training, see 'Automated Content Classification (ACC) Systems' (2023), a research report prepared by Winder.AI for Ofcom.

Removing violative content

4.3 **A central component of enforcement is the removal of violative items – that is, of content that is in breach of services' content standards.** Often, relevant processes are designed to reduce, rather than eliminate, the amount of violative content available and instances of users encountering this content. Specifically:

a) At a minimum, processes are generally designed to ensure that any item reported by end-users or other third parties as being potentially violative is reviewed and, if appropriate, removed.

b) Additionally, some service providers have developed AI-led systems to automatically detect potentially violative content. But because these systems are not error-free and their reliability varies across types of violation, in general they complement, rather than replace, human moderators.

c) Whether problematic content is initially reported by end-users or other humans or flagged as potentially violative by automated tools,[17] suspect content *may* remain available – albeit possibly demoted – until a human moderator confirms that it is violative.

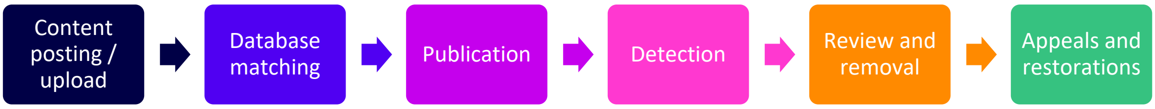
4.4 We discuss the relevant processes in more detail below.

A stylised process

4.5 Standards enforcement is generally a complex process that varies significantly across services. For ease of exposition, here we outline a simplified and stylised process that, in our view, is *broadly* reflective of the main approaches in use today on social and content-sharing services.[18] This process is summarised in Figure 3 and discussed below.

4.6 Our account is not necessarily fully representative of all cases or an exact reflection of any one service's process. It relates mainly to services where free-form content can be 'posted' and viewed by large numbers of users. For a discussion of some of the main variations, see page 20.

Figure 3: A stylised process for standards enforcement



[17] In this commentary we primarily refer to requests for review by users or 3rd party organisations as 'reports' and requests based on automated tools' assessments as 'flagging'. However, the terminology is often interchangeable in conversations about moderation, and its use can differ depending on the service. A piece of content being 'flagged' by an automated tool does not necessarily mean it has one or more 'flags' attached to it in the way it could have one or more user 'reports' – an automated assessment may be more about prioritisation (which we discuss in this paper) rather than a binary instance of flagged/not flagged. Note that the Digital Trust and Safety Partnership, an industry-led initiative looking to foster best practice, has released a glossary of definitions. While not prescriptive or legally-binding, this seeks to set out common terminology. In this resource the relevant definition is under the term 'flagging'.

[18] We have not focused on search engines in this paper, but we note that they also engage in activities relating to content moderation.

Content posting / upload

4.7 **Initially, a user posts a content item** – e.g. a video, an audio file, a picture or a piece of text. This may be an original item produced by the user posting; a 'repost' or 'sharing' of an item previously posted by another user; a link to an item found outside the service; or a comment responding to another user's post.

4.8 **Services may have restrictions on who can post content**. In some cases any registered user may be able to post – but registration itself normally requires users to meet certain criteria. In some cases, for example some adult services, only users who have undergone some form of ID verification and/or age assurance may be able to post. Services may also use other measures at the point of upload but before publication, such as interstitial prompts suggesting users may wish to reconsider posting in certain circumstances.

Database matching

4.9 **Once content is posted, automated systems may run a range of tests against submitted items**. These typically include checking whether the content posted is a copy of a previously known violative item, in which case publication would be prevented.

4.10 **This process is particularly relevant for catching known or close matches of illegal content, such as CSAM**. Shared industry databases containing identified instances of this content play an important role in this context.[19]

Publication

4.11 **Content that passes the matching stage will typically be published, i.e. made available to other users.** Depending on the service involved, the size of the potential audience of a published item may depend on user preferences (e.g. a user may specify that a post should be visible only to 'friends') and on where on a service an item is posted. In some cases, items may be available to all registered users of a service, or to all web users including those not registered with a service.

Detection

4.12 Once content is published (and, in some cases, possibly before publication – see below), **content may be identified as *potentially* being in breach of content standards**, not because it is a copy of previously known violative content, but as a result of:

a) **Automatic content classifiers**, typically using artificial intelligence (AI) technology, identifying an item as being *likely* to be violative of one or more policies. The development of content classifiers is an area of active innovation, with new approaches regularly being developed. A range of approaches may be in use at any given time even within a single service.[20]

An important common aspect of automatic classifiers is that, in general, a classifier's flagging of an item as being potentially violative would typically come with an indication of the *degree of confidence* associated to this assessment. If classifiers deem an item is likely to be violative with a sufficiently high degree of confidence, in some cases this may

[19] Shared databases may consist of identifying metadata about the content, such as 'hashing' fingerprints, rather than copies of illegal content. Database sharing organisations include the Global Internet Forum to Counter Terrorism (GIFCT) and the US National Center for Missing and Exploited Children (NCMEC).

[20] See footnote 16 above.

be followed by the immediate, **automatic take-down** of the item.[21] Automatic removals may take place only seconds after publication, in effect preventing all or nearly all viewing of affected items, or possibly even prior to publication (on which more below). Automated removals are common for certain specific types of content including spam;[22] services may also rely on automatic removals when human moderator capacity is insufficient – for example at the beginning of the Covid-19 pandemic, when some service providers operated with reduced human moderation capacity.[23]

Items not removed automatically are classified as potentially violative and then await a review, to be conducted by a human moderator. Depending on the service and content involved, items may remain available until they are reviewed (however, such items may be shown less prominently to users – 'demotion', discussed below).

b) **End-users or third-party organisations may report content for review**. Services typically require reporting users to specify from a list of options which policy they believe was breached. Some service providers work with trusted third-party organisations and may give their reports priority in the moderation process.[24]

4.13 **As technology improves and classifiers become more reliable at identifying violative content, user reports of violation may become comparatively less important in harm reduction.** Some services, and for some harms, increasingly see user reports of potential violation as less reliable than those of AI-based classifiers. Currently the vast majority of take-downs among some of the larger services reportedly result from detection by AI-based classifiers.[25]

Review and removal

4.14 **Human moderators (who may be service provider staff or subcontractors) review content reported by users/third parties or flagged by an automated tool as potentially violative**. If moderators find that content is violative, it is removed; otherwise it remains available. Content removals may also trigger user-level enforcement against the user who uploaded the item, such as temporarily banning new posts or account suspensions. Moderators may escalate difficult decisions to more experienced colleagues or senior leaders; instances of illegal content may be referred to law enforcement.

[21] For instance, Meta has published overviews of the use of artificial intelligence technologies to detect violations and, in some cases, proactively deleting it before users report, or sending on for review, in blog posts last updated in January 2022, called 'How technology detects violations' and 'How enforcement technology works'.

[22] For example, Pinterest notes in its transparency report that 99% of Pin deactivations for spam in Q4 2022 were due to fully automated tools (which they define as machine-learning assisted tools using a combination of signals to identify and take action against content) – compared to 3% of deactivations for self-harm and harmful behaviour in the same period; LinkedIn notes in its equivalent report that 99.3% of spam in H2 2022 was 'stopped by automated defences'.

[23] See e.g. p 5 of Facebook's May 2020 Press Call; see also YouTube's post 'Protecting our extended workforce and the community' (2020).

[24] For instance, YouTube notes in its Help Centre that reports from Priority Flaggers on YouTube are prioritised for review because these Flaggers are considered to have a 'high degree of accuracy'.

[25] For instance, in Q1 2023, YouTube's transparency report stated that 93.7% of videos removed from YouTube were first detected through automated flagging. This does vary by violation type - for instance, across Microsoft's Xbox services 100% enforcements in H2 2022 for account tampering, piracy, phishing were reportedly the result of proactive technologies and processes, but the proportion for adult sexual content was 27.9%, according to Xbox's transparency report.

4.15 **In the case of human review, there may be a time-lag between content being identified as potentially violative and it being reviewed and removed by moderators,** as a consequence both of the volume of items being referred for moderators' attention (a volume that may fluctuate,[26] e.g. in response to news events) and human moderation capacity.

4.16 **How items should be prioritised is not a question with a straightforward answer**; for example, a simple first-in-first-out approach to tackling a 'backlog' might mean that a recently-posted, highly harmful 'viral' item attracting millions of viewers has to wait until all older reported items, including comparatively harmless items, have been reviewed. Thus, some service providers devote considerable thought to designing the systems that determine in what order items are sent to human content moderators. For example, Meta has disclosed[27] that its prioritisation criteria on Facebook and Instagram include:

a) the estimated likelihood that an item will be confirmed as violative, as discussed above – which may lead to more actually infringing items being reviewed and removed;
b) the 'virality' of an item (how quickly it is being shared) – so that moderators' efforts are focused on items that may be viewed by the largest volume of users; and
c) the severity of the potential harm in question – so that items more likely to lead to harm are reviewed earlier.

Appeals and restorations

4.17 **Appeals and restorations offer users an opportunity to challenge moderation decisions.** Users whose content was removed (or whose account was suspended) can normally submit an appeal, and may be given the opportunity to select a reason for the appeal or offer additional context or explanations. The content in question is then reviewed, and moderators may uphold the removal decision or restore the content (and/or account).

4.18 A service may also enact restorations without appeal, for example if it determines that an error in its systems resulted in excessive false positives and over-removal. Some services will also notify users or third parties who reported content found not to be violative, offering the option to appeal against that decision.

Key ways in which some services differ from our stylised account

The sequencing of automatic detection and publication can vary

4.19 **The sequence discussed above describes the AI-driven detection of previously-unseen violative content as taking place *after* publication**. The exact timing of this is not something we have looked into in detail, and it is possible that in some or even in most cases it may happen prior to publication. We note however that, provided that automatic detection takes place only seconds after upload (which we understand is generally, but not always, the case), whether it happens shortly before or after upload should make only minimal or no difference to user outcomes.[28]

[26] For example, Meta's Oversight Board disclosed in 2022 that by late 2021 the company was "performing about 100 million enforcement attempts on content every day".

[27] See Meta's published information on this at 'How Meta prioritises content for review' (updated 2022) and 'How we review content' (2020).

[28] If AI-flagged content is left 'live' pending review by moderators, removal would not happen until after review, which may be minutes or hours after detection, and in this context the effect of a few seconds' delay

Communities and volunteer moderators play an important role on some services

4.20 **Some services include functionality allowing users to form communities whose members can post content within a certain area of the service** – key examples include Facebook's 'groups' and Reddit's 'subreddits'. Posts shared within communities may or may not be visible to service users who are not members of a community.

4.21 **Communities may rely on volunteer users to moderate content according to a community's own rules, which complement services' overall policies.** On Reddit, user reports of content suspected of violating the service's overall content standards (which take precedence over communities' own rules) go to both the service's own safety employees (known as 'Reddit Admins') and volunteer subreddit moderators, who may each take certain actions, depending on the violation. This contrasts with reported violations of individual community rules, which are handled solely by the relevant volunteer moderators of that community.[29]

4.22 User moderators also play an important role in content moderation on services such as Discord, Twitch, Nextdoor, Mastodon and Wikipedia.

Services providing private (one-to-one) communications not widely visible to other users may rely on different approaches to reducing harm

4.23 **Some services may not offer one-to-many functionalities whereby a user can 'post' content that may be viewed by a wide range of other users**, for example through 'news feeds', lists of recommended content, or other similar facilities. This includes dating sites as well as user-to-user messaging apps (standalone or as in-app products within larger services). In these cases, content may be specific to private conversations and intended for only one or a few recipients, often with an expectation that content will reach its recipient quickly (e.g. through email notifications). In these contexts, content moderation processes which aim at reducing the *number* of users that see a violative item, like the process we have outlined, may be less relevant.

4.24 **Instead, in such contexts harm reduction may rely more on tools** such as user identification at sign-up; reporting of non-compliant users; discouraging certain behaviours such as reposting viral content which the user has not read; and user-led tools to prevent viewing of undesired content (e.g. warnings about messages from people with whom the user is not "connected" and sensitive content warnings). For example, Bumble offers users of its different services the ability to filter out nude images and bad language in incoming messages.

Smaller service providers have more limited access to the technology required to automatically detect potentially violative content

4.25 **The development of AI-led classifiers is a costly endeavour** requiring cutting-edge technical talent, abundant data used in the training of these systems, and other resources that may be beyond the means of smaller service providers. Such providers may be able to source some of these capabilities from third-party vendors, or from peers who may offer this capability

in removal, resulting from a few seconds delay in detection, may be minimal. Alternatively, if automatic detection is immediately followed by automatic removal, an affected item would be 'live' for only a few seconds. However, we also understand that the automatic detection of certain kinds of violation may sometimes be less fast, given the computational resources required.

[29] For more on Reddit's reporting processes and how Admins prioritise reports, see 'On reports, how we process them and the terseness of "the admins"' (2018).

commercially. However, this may require resources to tailor third-party tools to smaller services' specific content policies and functionalities. This is an area of active innovation which we expect will continue to evolve over the coming years.

The human screening of all content is done in certain cases, but it is uncommon

4.26 **Requiring human moderators to review all content prior to publication – 'screening' – is a relatively uncommon practice.** However, across-the-board screening *is* used by certain service providers; for example, in 2022 adult service PornHub disclosed that it screened all content,[30] as does children's app Lego Life.[31] Other service providers may screen content more selectively, such as e.g. if it is to be featured on certain parts of the service.

Applying non-takedown measures

4.27 As noted earlier, **some service providers take measures to reduce the visibility of various types of potentially problematic content that is not necessarily violative**. This could be 'borderline-violative' content; content displaying 'signals' of being potentially violative that has not yet been reviewed and confirmed as such; or other non-violative content deemed undesirable by the service provider for other reasons (e.g. misinformation, sensationalistic content or click-bait). Additionally, some service providers may target content that they expect users may not like or value (e.g. based on 'down votes', 'dislikes' or surveys[32]). The relevant types of intervention are summarised in Figure 4 below:

Figure 4: Treatment of violative and other content – a summary, not necessarily applicable to all services

Actions	Violative content	Borderline-violative	Other problematic
Removals	Yes – after review	No	No
Demotions and / or other measures	Yes – if there are signals that content may be violative. Once reviewed, content found to be violative is typically removed; borderline content may remain demoted.		Yes

4.28 We now outline the main types of non-takedown *measures* employed:

Demotion (or 'down-ranking')

4.29 **Service providers may make content appear less frequently or prominently in users' news feeds, in services' home pages and in lists of recommended content** – thereby reducing the likelihood of affected content being seen. There are several variants of this, depending on the service involved: [33]

[30] As per PornHub's 2021 transparency report: "all content is reviewed upon upload by our trained staff of moderators before it is ever made live on our site".

[31] As noted in the Help section of Lego.com, about the Lego Life app: "We have a team that's moderating all content before it's posted for everyone to see".

[32] For instance, Reddit's voting system impacts on what content is displayed across the platform, as briefly explained in the Help section of the site.

[33] Services also take measures to increase the visibility other, higher-quality or trusted content. This may also result, indirectly, in reducing the visibility of content deemed problematic.

a) Services may make an item less visible by ensuring it only appears once users have scrolled through many other items. *In extremis*, an item may be altogether excluded from all lists of content – in effect ensuring that it is seen only by users expressly looking for it (e.g. through a URL or a very specific search query).[34] We understand that for at least some services the demotion or exclusion of potentially problematic items from feeds or lists of content may differ from user to user.

b) An item may appear in news feeds and other lists only for users who are members of a specific user group or community to which the content was posted.

c) An item may be featured in feeds and lists only for users whose previous behaviour suggests that they might like to see the content in question. However, if an item is deemed entirely ineligible for feeds, lists and other promotion (as per (a) above) it may not be offered even to these users in such contexts.

4.30 **Generally, demotion measures are applied algorithmically** – that is, the question of whether they should be applied to a given item is decided automatically, based on data about the content in question (among other things). Demotion measures may be applied to items that otherwise might be particularly popular or 'viral', for reasons unrelated to the service (for example a conspiracy theory post might 'go viral' due simply to many users choosing to share the item with their friends).

4.31 Separately, **services may algorithmically increase, rather than reduce, the visibility of certain items for a variety of reasons** – e.g. to drive engagement, 'meaningful interactions', user satisfaction as measured through surveys, or the viewing of high-quality content.[35] The factors driving prominence, and the way these are combined, vary across services and have evolved over time. [36]

4.32 **It is possible that an item may simultaneously, and inconclusively, exhibit (a) characteristics that, in isolation, might lead to promotion; and (b) characteristics that, in isolation, might lead to demotion.** For at least one of the service providers we spoke to, how such content should be handled is a difficult question to which significant attention may be devoted, potentially including extensive internal review and calibration among multiple expert teams.[37]

Overlays, 'interstitials', 'panels', blurring and labels

4.33 **Content may be accompanied by, or covered with, messages noting that the content in question may be disturbing, may contain disputed claims or may be otherwise sensitive.** Services may offer links to supporting organisations and resources (such as next to posts related to suicide or Covid-19), including other materials within their own sites. Overlays (or

[34] For example, see YouTube's blog post, 'Inside Responsibility: What's next on our misinfo efforts' (2022): "We've overhauled our recommendation systems to lower consumption of borderline content that comes from our recommendations significantly below 1%. But even if we aren't recommending a certain borderline video, it may still get views through other websites that link to or embed a YouTube video".

[35] On the promotion of content, see YouTube's blog post, 'The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation' (2019).

[36] See, for example, YouTube's blog post, 'On YouTube's recommendation system' (2021), a Facebook Business article, 'News Feed FYI: Bringing people closer together' (2018), Instagram's blog post, 'Instagram Ranking Explained' (2023) and Meta's AI "systems cards" (updated 2023).

[37] For a discussion of how borderline-violative content may be particularly engaging (organically, without the aid of algorithmic promotion), see a post by Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement' (2018).

‘interstitials’[38]) and blurring may cover an entire piece of content and require the user to click through. A label may provide a warning or additional context (e.g. for disputed claims). Labels may be added by the service, or in some cases content creators themselves have the ability to label their own content e.g. NSFW (‘not suitable for work’) or ‘trigger warning’.

4.34 **A service may also provide in-app notifications for the creator of content that is covered or labelled.** This offers transparency about the interventions being applied, which can then be appealed if the user believes it is incorrect. In addition, by informing the creator that their content may be problematic (and may be getting less engagement), a notification may encourage them to delete, edit or clarify the original.

Other interventions

4.35 **Service providers may also use other, non-takedown measures,** including:

a) Prompts encouraging users to pause and think before posting content that appears potentially problematic;[39]
b) Restrictions on posting, sharing and forwarding, for example if posts contain specific links, if a user has shared a large number of items, or if an item has been forwarded through a long chain of onward ‘reposts’; and
c) Age-gating to restrict access to certain content or versions of a service.

Such measures may be said to reflect a ‘safety by design’ approach to safety.

[38] This is a term used in the ad industry to refer to advertisements that cover the whole of the host app, typically at points between activities or between pieces of content. In the case of Trust and Safety interventions, they typically cover only the content affected.

[39] For example, X may prompt a user before posting a tweet that may include hateful content. The effectiveness of this and similar techniques has been a question of some interest in the expert community. X itself (then called Twitter) contributed to published research investigating this intervention (2022).

Part II: Our reflections

5. Decisions and trade-offs in content moderation

5.1 **In general, content moderation processes are designed to limit the viewing of violative content, rather than to guarantee that no harmful or violative content can be accessed.** For example, as we saw in section 4:

a) Typically content moderators do not see all items posted, but rather only those items that have been referred by users or AI classifiers – neither of which are guaranteed to catch all violative items. Violative items that go undetected by end users and classifiers may never be seen by a moderator or removed.

b) Items referred to moderators may remain available until they are reviewed. This means that for the period between publication and review, harmful content may be accessed by users.

c) Although services may rely on automated systems to demote and possibly remove certain types of potentially violative items before these are reviewed by moderators, these systems may fail to catch some items; moreover, content demotions generally do not aim to prevent all access to content, but only to make content less widely visible (so that, e.g., content is offered only to users with certain interests).

d) Even if all content were to be screened by moderators in a timely fashion a certain amount of harm may still result from human error, ambiguities in content standards (e.g. around borderline cases), or from content standards not covering every situation where there is even a remote chance of causing harm.

5.2 **These limitations reflect decisions in process design, and these decisions may entail trade-offs which may or may not be explicit**. For example, a decision to allow items reported as potentially violative to remain 'live' until a moderator sees them[40] may lead to fewer non-violative items being removed erroneously than if such items are automatically removed pending review – but potentially with the result that violative items remain available for longer, and are viewed by more users, than otherwise. Thus, this decision may be seen as entailing a trade-off between enabling users to express themselves freely on the service, and keeping users safe. Process design choices may also have impacts on service usability, user satisfaction and on services' operating costs incurred by service providers. Trade-offs like these may be unavoidable.

5.3 **Within the constraints set by the overall structure of their processes, service providers possess a range of levers that can reduce harm**. For example, a service relying mainly on human moderators to remove content may be able to reduce – through automated means – the visibility of content awaiting review; it may be able to reduce 'turnaround time' between content upload and review (and removal) by hiring more moderators; or it may choose to allow machines to remove specific types of potentially high-risk content, freeing up moderators' time or other types. However, again each of these levers entails costs and trade-offs, and none offers a silver bullet.

[40] For example, YouTube notes in its blog, 'On policy development at YouTube' (2022): "For most categories of potentially violative content on YouTube, a [classifier] simply flags content to a content moderator for review before any action may be taken".

5.4 We explore some of these levers and trade-offs in this section. Following the structure of Part I of this report, we first cover decisions made within policy-setting and then decisions made in process design and management (although this distinction may not be clear-cut). While this section draws on discussions with service providers, all observations here are our own. Service providers may not share our views regarding what the key decisions and trade-offs involved are, or indeed whether the decisions and trade-offs we discuss are involved at all. Service providers were not offered an opportunity to provide comments on this section prior to publication.

Decisions in policy-setting

How broadly or narrowly to define violative content

5.5 **A central decision that service providers must make in content moderation is what types of content to ban, and how to reflect this in clear and enforceable content standards.** In some cases, services may ban broad content categories that may capture both harmful and potentially harmless content. For example, a service might choose to ban all nudity (including non-pornographic, consensual adult nudity) as a way to minimise the possibility that harmful types of nudity might go undetected, and/or for commercial reasons (e.g. advertisers' brand safety).

5.6 **However, broad bans may be controversial on services or in contexts where users may expect a degree of unrestricted expression**; for example, historically Facebook and Instagram's policies on nudity have attracted criticism for resulting in the removal of content with cultural or educational value. [41] Services may seek to address such concerns through policies aimed at ensuring that otherwise violative content may remain available on grounds of (for example) educational value, relevance to public debate or newsworthiness.[42]

5.7 **Alternatively, services may opt to more narrowly target content that they consider harmful or unacceptable.** This may be favoured by services where unrestricted expression is an important aspect of the customer proposition. For example, X, formerly Twitter, permits "adult nudity and sexual behaviour content" that is consensually produced, is not violent, and does not contain "gratuitous gore".[43] However, the use of narrow bans may require services to spell out where the boundary of acceptability lies, and to deal with 'borderline-violative' content whose violative/non-violative status may be unclear.

[41] In 2016 Facebook received criticism for removing a 1972 image of a young, nude girl in Vietnam fleeing an attack on her village, under their policy to ban all imagery of child nudity. They subsequently restored the image because of its historical importance (for media coverage see, for instance, Reuters 'Facebook reinstates Vietnam photo after outcry over censorship' (2016)). Currently some services have nuanced policies that allow imagery (maybe with additional labelling) of adult nudity in medical or educational contexts, and child nudity in news about war or similar. In general, concerns around censorship still frequently feature in discussions around how far content standards should go.

[42] See, for example, Meta's article 'Our approach to newsworthy content' (updated 2023) and X's Help Centre piece, 'About public-interest exceptions on Twitter'.

[43] Content passing these tests may be covered by interstitials and is only available to adult users. See, for instance, X's Sensitive media policy.

How aggressively to apply demotions and other non-removal measures

5.8 **The use of demotions, labels or overlays can be seen as reflecting a trade-off between the need to protect users from harm and the desire to allow users to express themselves on the service** – for example by ensuring that content that may upset some users is only seen by those who actively choose to access it, without removing it outright**.** But here again how far such measures should be deployed is a difficult question, as it may not be practicable to simply decide that all items that are potentially violative, borderline violative, sensationalistic or of low-value should always be demoted or labelled.

5.9 **If items have not yet have been reviewed by a moderator, an automated enforcement system may only 'know' that an item *may* merit demotion based on an AI-assigned *probability*** – and many harmless items may also be judged to have at least *some* probability of harm. Thus, in deciding how aggressively to apply non-demotion measures, service providers must weigh the benefits in harm reduction against the risk of unintentionally demoting more and more harmless items (we discuss 'false positives' further below).

5.10 **Service providers must also decide how to handle conflicting signals favouring promotion and demotion of the same item**. As noted in section 4 above, enforcement systems may deem a content item to likely merit down-ranking, as above, even if at the same time systems also deem the same item to be likely to elicit engagement, user satisfaction, or other outcomes sought by the service. In designing how systems should handle such cases, 'erring on the side of demotion' may not be practicable since, as noted above, many (if not most) unproblematic items may exhibit demotion-triggering signals at least to a small degree, while there might be strong confidence that some such items will lead to high satisfaction among certain users. Whether such items are ultimately promoted or demoted, and how much, is a question that service providers implicitly or explicitly answer through the configuration of their systems.

Decisions in enforcement process design

How to prioritise items awaiting moderation

5.11 **As we noted in the previous section, commonly there is a time-lag between the moment that an item is first detected as being potentially violative (either by users or AI classifiers) and it being seen by a human moderator**. This partly reflects the fact that human moderator pools, however large, have finite capacity, and the amount of potentially violative content posted every day is not only vast but also liable to sudden and unexpected increases, for example in response to news events. In addition, the speed with which a service provider can review content relating to specific harms or from specific countries where a 'spike' of problematic content may originate could depend on the number of staff within the overall pool that have relevant expertise, cultural knowledge or language abilities.

5.12 **There may not be any meaningful sense in which moderators might aim to 'clear the backlog' of content awaiting review.** Even if at any point moderators were to finish reviewing all items reported by users plus all high-confidence AI-based referrals, there may well be many more items, classified with lower degrees of confidence but which it may be beneficial to also review, as some of these items may indeed turn out to be violative. Once moderators complete reviewing these, they could turn to a yet larger pool of items classified with yet lower confidence, and so on.

5.13 **In this context, systems can be designed to prioritise the review of suspect items using a range of criteria, each of which leads to different impacts on harm, user expression or commercial outcomes**. For example, if priority is given to items *most likely to be confirmed* as violative by a moderator, this may maximise the number of violative items removed; if priority is given to potentially violative items *likely to be most harmful*, this may minimise the potential for more serious harm; if priority is given to potentially violative items attracting *most viewers*, this may minimise the number of viewers affected; and if priority is given to items that have been waiting the longest for review, this might help ensure that items reported by users are reviewed within a set period of time.[44]

5.14 This entails some difficult questions. For example, suppose that at a given point in time a service is aware of two potentially violative items awaiting review (alongside thousands of other items):

a) Item A, suspected with high confidence of being highly harmful (e.g. particularly abusive hateful speech), but being viewed by only 100 users per hour;
b) Item B, suspected of being only moderately harmful, with less confidence, but receiving 10,000 views per hour, and rising – as the item has 'gone viral'.

While both items may eventually be reviewed by moderators, which item gets reviewed *first* may have a significant impact on safety outcomes — option (a) means that relatively few people are protected from significant harm, while option (b) means that a large number of people are protected from unconfirmed mild harm. As noted in section 4, in practice services may use a combination of criteria like these.

5.15 Dilemmas like the above arise partly as a consequence of the fact that human moderators cannot review every item immediately upon it being uploaded. Service providers may be able to mitigate this by hiring more moderators and/or relying on machines to moderate content; however, both of these involve further decisions and trade-offs, as we see next.

How much human moderation capacity to have

5.16 **By growing its human moderator capacity a service provider can increase the number of items it can review per hour, in turn potentially leading to shorter delays between content detection and removal, and/or to a greater proportion of posted violative items being reviewed.** For example, a service could aim to ensure that it can review, within any 24-hour period, all[45] content that end-users report as being potentially violative; or it could aim to do so within each hourly period; or it could aim to do all this as well as processing all AI-flagged content where classifiers have, say, a 80% confidence that content is violative; or variants thereof. Moreover, in some cases increased moderator capacity, or enhanced moderator

[44] The Network Enforcement Act (2017) in Germany (known as 'NetzDG') requires that services take down or block access to obviously illegal content within 24 hours of receiving a complaint and all other illegal content within 7 days of receiving the complaint (with some exceptions for this time limit). The services must report publicly on their handling of complaints, and several services have dedicated pages for these transparency reports e.g. YouTube's 'Removals under the Network Enforcement Law' and Twitch's NetzDG Transaprency Report (H1 2023).

[45] More precisely and due to given ebbs and flows in demand, a service might aim to meet this target, say, on 95% of all days (or, at a higher cost, 99% of all days).

training, might also allow for more careful reviewing of difficult cases, thereby potentially reducing error.

5.17 **While increased or more specialist capacity may lead to better safety outcomes, of course this comes at a cost**, and currently each service provider must decide how much to spend on its moderator workforce. Where moderation processes rely on moderators seeing only a small proportion of items posted, service providers may ask themselves whether it is worthwhile to spend an extra £x million, or an extra y% of their revenue, to achieve a further z% reduction in the number of users who see violative content (or in some other metric of harm). This question may be particularly relevant if the degree by which harm is reduced diminishes with each additional unit of additional investment; this might be the case, for instance, if a service's prioritisation processes ensure that moderators review the most harmful and most viral content first, so that increasing the size of the moderator pool may result in individual moderators reviewing fewer harmful and/or viral items than before the increase, on average (even if the moderation team's overall throughput increases) .

5.18 **We note, finally, that in deciding how much and what types of moderation capacity to have, service providers may also need to consider whether a greater impact may be achievable by investing in technology**; how long and this might take to materialise; how likely the service provider thinks it is that technology will bring significant benefits; and ultimately what the best mix may be of spend on current human moderator capacity and on long-term investment in technology.

To what extent to rely on algorithms to make automatic removal decisions

5.19 **A further way to mitigate problems associated with limited moderation capacity may be to allow machines to remove more content without human involvement.** Often, AI-driven decisions can be reached very quickly. But classifiers are not error-proof: they may fail to detect some violative items ('false negatives'), particularly for certain types of violation, such as harassment, where assessing whether content is violative requires an understanding of context and nuance; and they may also wrongly remove items that are not violative ('false positives').

5.20 **Often service providers can adjust classifiers' settings to reduce false negatives at the cost of incurring more false positives, and vice versa**. When deciding whether and to what extent to rely on automatic moderation, providers must often take a view not only on how many 'false negatives' and 'false positives' to accept, but also on whether an increase in one is an acceptable trade-off for a reduction in the other.[46]

Whether and when to implement pre-moderation / screening

5.21 **We noted in Section 4 that it is rare for a service provider to require human reviewers to screen all content prior to publication.** If a service were to have enough moderators to review every item posted, *and* it were to require that content only be published after being

[46] In the technical literature, the need to make trade-offs between a high rate of 'false positives' and a high rate of 'false negatives' is referred to as the tension between 'recall' and 'precision' rates (with high a 'recall' rate being associated to few false negatives, and a high 'precision' rate to few false positives). For more on the technical processes underpinning automated classifier systems, see our reports 'Use of AI in online content moderation' (2019) and 'Automated Content Classification (ACC) systems' (2023).

vetted by a moderator, then exposure to violative content might be reduced significantly, with remaining viewing due largely to human error in enforcement or to a failure to capture harmfulness in content standards. However, in addition to the cost-related considerations discussed above (which may be prohibitive), the introduction of general screening would amount to a major departure from current practices and might be perceived negatively by some users.

5.22 **This type of pre-publication screening is likely only to be relevant for design considerations in certain contexts** where the risks of severe harm (such as from illegal activities and content) is high and the amount of new content being uploaded is low.

6. Measuring the performance of content moderation efforts

6.1 **Inasmuch as content moderation can only reduce but not entirely eliminate harm, its success is a matter of the *degree* by which harm is reduced. A key question that then arises is: how might this be measured?**

6.2 **In recent years several service providers have started publishing certain quantitative metrics that – by their own accounts – play a central role in how providers assess their own performance in content moderation.** Key examples include Meta's 'prevalence' ("the number we hold ourselves accountable to"[47]) and YouTube's 'violative view rate' ("the primary metric [we use] to measure our responsibility work"[48]), both of which are estimates of how often users are exposed to violative content. For an overview of relevant metrics see Box 3 below.

Box 3: Current publicly-available exposure metrics and their evolution

Many service providers routinely gather and publish a wide range of quantitative data relevant to content moderation.[49] Most of the available data focuses on 'supply-side' factors – i.e. on the violative content detected within services and how services deal with this. For example,

- **Many service providers routinely publish data on the number of removals** carried out over a specific period, broken down by type of violation and by how content was detected (e.g. flagged as potentially violative by classifiers or reported by end-users, official bodies or partner organisations). Examples include Pinterest's 'Total Pins deactivated', TikTok's 'videos removed' (which also includes breakdowns such as the proportion of videos removed before being reported by users) and YouTube's 'videos removed by source of detection'. Facebook and Instagram's 'content actioned' covers both removals and some non-removals actions, not broken down, and as with TikTok is supplemented by data showing the proportion of items actioned before being reported by users.
- Some service providers publish estimates of the proportion of available content that has been removed. This metric does not account for the number of views that content gets, but could be an (albeit limited) indication of the potential

[47] See the report accompanying the Facebook transparency report for Q4 2017 to Q1 2018, 'Understanding the Facebook Community Standards Enforcement Report'. Meta CEO Mark Zuckerberg has also written: "We think prevalence should be the industry standard metric for measuring how platforms manage harmful content." See also Meta's Newsroom post 'Measuring Prevalence of Violating Content on Facebook' (2019).
[48] YouTube's blog post, 'Building greater transparency and accountability with the Violative View Rate' (2021). This also states: "Our teams started tracking this back in 2017, and across the company it's the primary metric used to measure our responsibility work."
[49] The examples of transparency reports which we reference here and elsewhere in this paper are those published by Meta (Facebook and Instagram), YouTube, Reddit, Pinterest, TikTok, BitChute, Snap, X, PornHub and Xbox (note this is not an exhaustive list of all transparency reports available).

availability of violative content. For instance, BitChute has disclosed that 0.0003% of all its 'in-scope content' was removed on the grounds of harassment in Q1 2023. Meanwhile Reddit has published a number of related datapoints that together suggest that around 4% of the content created in FY 2022 was subsequently removed.[50]

In recent years, a number of service providers have attached particular importance to metrics that focus on end users' exposure to violative content – rather than on how much violative content is available or removed. Specifically, providers have increasingly focused on metrics that estimate how many times users view violative content, regardless of whether the content was subsequently detected and removed**.** Examples of such metrics in public transparency reporting include

- Facebook/Instagram's 'prevalence' (for which they provide several splits by type of violation). For instance, Facebook reports that around 0.08% of content views (8 in every 10,000) are of bullying and harassment material (Q2 2023).
- YouTube's 'violative view rate', which is aggregated for all harms excluding spam. YouTube reports that 0.08-0.10% of all content views (8-10 in every 10,000) are of violative content (Q1 2023).
- Snap's 'violative view rate', which (like YouTube's) is aggregated for all harms. Snap reports that 0.03% of all content views (3 in every 10,000) are of violative content (H2 2022).
- Reddit occasionally releases estimated of viewing of hateful speech.[51]

As a variant of these, some service providers publish estimates of the amount of viewing accrued by items that were subsequently removed. For example:

- X, formerly Twitter, has previously published a metric called 'impressions of violative Tweets': X reports that 71% of removed posts (formerly Tweets) had fewer than 100 impressions, and that less than 0.1% of all impressions (fewer than 1 in every 1000) were of violative posts (H2 2021).[52]
- Pinterest's 'reach of deactivated Pins': Pinterest reports that 65% of Pins deactivated for 'harassment and criticism' were not viewed (Q4 2022).

6.3 **We welcome service providers' contributions to a richer public understanding** by drawing attention to questions around what outcomes should matter in content moderation and how these can be measured; their proposals around a concrete, measurable set of metrics;

[50] Reddit publishes figures for total content created (posts, comments, private messages and chats), and total content removed by both community moderators and admins. Note that content authors can also remove their own content – this has been excluded from the calculation. It should also be recognised that removals may relate to content created in a previous year, so this is an estimate only.

[51] Reddit, 'Understanding hate on Reddit, and the impact of our new policy' (included data on the prevalence of hateful content) (2020); 'Q2 Safety & Security report' (included data on the prevalence of Holocaust denial) (2021).

[52] As of September 2023, X's transparency centre shows data up to and including July to December 2021. When accessed in September 2023, the page is not showing all the charts as originally published, but the information about impressions is still available. The most recent update from X pertaining to transparency reporting was published in April 2023, relating to the period January to June (H1) 2022, but this does not include the impressions metric.

and their decisions to publish corresponding data. Ofcom believes that greater transparency may empower and inform several parties including users, researchers and investors; hold industry accountable; and incentivise improvements.[53]

6.4 While the available data is certainly valuable, certain voices in the expert community have called for service providers to publish more granular data, for example showing to what extent exposure to violative content may be concentrated among small groups of users.[54]

6.5 We note also that available metrics offer limited insight on how exposure to violative content may vary across age groups or other demographics and protected characteristics, countries and (in some cases) type of violation, content or service functionalities; or on how frequently users see *borderline-violative* content or other content targeted by visibility-reducing measures.

6.6 **In our view, public understanding of trust and safety systems and their effectiveness would benefit from information on such aspects**. We look forward to engaging with service providers, online safety regulators in other jurisdictions, and other stakeholders in the coming years on these issues in the context of Ofcom's new role as the UK's online safety regulator.

[53] Ofcom has presented its regulatory strategy for transparency and metrics in the UK context in the following article: Harling, A-S., Henesy, D., and Simmance, E. 2023. 'Transparency Reporting: The UK Regulatory Perspective' *Journal of Online Trust and Safety*, p.1-8.

[54] For example, the Integrity Institute's 'Metrics & Transparency. Data and Datasets to Track Harms, Design, and Process on Social Media Platforms' (2021), 'Ranking and Design Transparency' (2021), and 'Shining a Light on Platform Transparency Best Practices' (2023); and Samidh Chakrabarti's comments via X, on "p99 prevalence" (2021). Others have suggested further avenues for measurement, such as of *positive* engagement and experiences, as proposed by Ravi Iyer in the post 'Content Moderation is a Dead End' (2022).