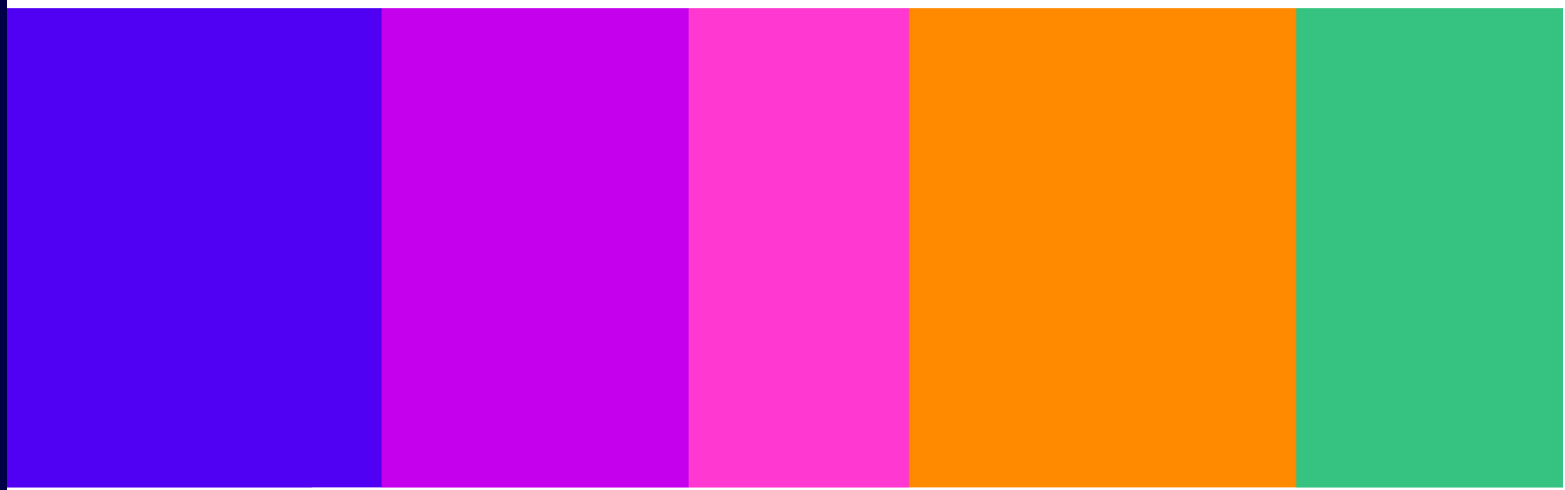
# [Final Draft] Guidance on appropriate proportion of human review for hash matching

## Statement

Issue date: likely February 2027. This is a DRAFT published on 16 November 2026 to be issued when the associated Codes of Practice are issued.

For more information on this publication, please visit ofcom.org.uk



# Contents

## Section

1. About this guidance ..... 3

# 1. About this guidance

**About this guidance**

This is our guidance to assist providers of regulated user-to-user services and search services (where applicable) in considering the appropriate proportion of content subject to human review when implementing hash matching processes for the purposes of detecting intimate image abuse, terrorism content and child sexual abuse material (CSAM).

## The measures

1.1 Measure ICU C9 of the Illegal Content Codes of Practice for user-to-user services recommends that certain service providers should deploy hash matching technology to detect image-based CSAM (the CSAM Hash Matching Measures).

1.2 Measure ICU C13 of the Illegal Content Codes of Practice for user-to-user services recommends that certain service providers should deploy hash matching technology to detect terrorism content (the Terrorism Content Hash Matching Measures).

1.3 Measure ICU C14 of the Illegal Content Codes of Practice for user-to-user services and measure ICS C8 of the Illegal Content Codes of Practice for search services recommend that certain service providers should deploy hash matching technology to detect intimate image abuse content (the IIA Hash Matching Measures).

1.4 This is our guidance to assist providers of regulated user-to-user services and search services (where applicable) in considering the appropriate proportion of detected content subject to human review for the purposes of these measures.

1.5 Nothing in this guidance, insofar as it relates to the processing of personal data, either supersedes or derogates from the requirements of applicable data protection legislation. Providers should continue to refer to the relevant guidance issued by the Information Commissioner's Office to ensure they are complying with their data protection obligations.

## Human review

1.6 Human oversight is an essential and well-recognised safeguard in ensuring that the overall outputs of content moderation processes which use automated technology are accurate, fair and contextually appropriate. Human review supports accountability, protects user rights and enables corrective action where detection errors or edge cases arise. When technological outputs are explainable (meaning, it is possible to understand how and why they were generated) human reviewers are better able to assess and make informed decisions.

1.7 For the Measures, there are two purposes for human review:

a) Moderation: service providers may need to use human review to determine:

i) whether the content is an intimate image (for example, if there is a first positive match with an unverified hash and the provider does not use automated technology to determine whether the image is intimate);
ii) whether the content is terrorism content (for example, considered in its context); or
iii) whether the content is known image-based CSAM; and

b) Quality assurance: using human moderators for quality assurance and monitoring of automated technology.

1.8 Providers should allocate human review resources on the basis of documented performance evidence and should review and adjust this allocation regularly to reflect observed accuracy, error patterns and risks of harm.

## Human review for moderation purposes

1.9 The measures say that service providers should ensure human moderators review and assess an appropriate proportion of detected content.

1.10 In doing so, service providers need to have regard to the level of assurance the provider has in the detection outcomes produced by the technology and any associated systems and processes. This assurance can be determined by ongoing monitoring, evaluation and quality assurance (including human quality assurance) of the performance of the technology. For example, where perceptual hash matching is configured to prioritise recall, a greater proportion of human review is likely to be appropriate to mitigate the increased risk of false positives, unless other technologies are used to verify the content.[1] The level of assurance reflects the degree of confidence that a provider can reasonably place in the system as a whole.

1.11 For content moderation and search moderation (where applicable to the IIA Hash Matching Measures) purposes, they should also have regard to:

a) the potential severity of the harm to those depicted in or encountering intimate image abuse content, terrorism content or known image-based CSAM; and
b) the overall impact of an incorrect decision that the content is illegal content, on the user who generated, uploaded or shared the content (for user-to-user services) or the interested person (for search services, where applicable). In relation to known image-based CSAM, this includes the importance of minimising the reporting of false positives to the National Crime Agency (NCA) or a foreign agency (within the meaning of Chapter 2 of Part 4 of the Act).

1.12 The impacts of an incorrect decision that the content is illegal content include, for example:

a) Restrictions on a user's access to the service, or their ability to use the service; and
b) An adverse impact on the user or interested person's financial remuneration payable in connection with, or generated by, the detected content (such as advertising revenue, subscriptions, commission or other monetisation arrangements), where known.

1.13 In considering detected content which may have a high impact on the user, the provider should also consider the extent to which and how promptly the relevant content moderation actions can be reversed.

1.14 In the rest of this guidance, we draw out some issues which are specific to the kinds of harm concerned.

### Terrorism content hash matching

1.15 In practice, we expect that the level of assurance in the detection outcomes produced by the hash matching technology will be low for the majority of the terrorism offences. This is because most terrorism offences are contextual, requiring providers to consider the context

[1] Recall is the proportion of true positive cases that are correctly identified.

of the content, including any relevant intent of the posting user, when determining whether the content is illegal. At the time of issue of this guidance, Ofcom was not aware of any available technology that could reliably make such contextual assessments without human input. Therefore, human review is required for most hash matches. A service provider should not rely solely on automated review in relation to content requiring contextual assessment unless it is confident that the automated review can correctly determine the relevant context.

1.16 We consider that a provider may be able to make an illegal content judgement without further contextual assessment where the content is known to amount to:

a) the offence of collection of information likely to be of use to a terrorist; or
b) instruction or training in the making or use of radioactive material, weapons designed or adapted for the discharge of any radioactive material, explosives, chemical, biological or nuclear weapons. This includes blueprints for printing 3D firearms.

1.17 Human review may still be needed of first hash matches of content falling within these offences, as a hash match alone cannot determine the specific priority terrorism offence engaged. However, once a provider has assessed content detected as a match to a hash and determined the relevant offence(s), subsequent matches of content to that hash may be sufficient to make an illegal content judgment without further review, provided the perceptual matching distance thresholds are narrow enough that the provider can have a high degree of assurance that the detected image is a match of the original item.

## CSAM hash matching

1.18 Context is not relevant to determining whether matched content is CSAM. Content included in the most widely used CSAM hash databases is subject to verification before inclusion, and perceptual hash matching technologies can be configured to achieve very high levels of accuracy. If hash matching is implemented in accordance with the safeguards we have set out in ICU C9, it should give rise to relatively small numbers of false positives.

1.19 If the prevalence of image-based CSAM on a service is very low, even highly accurate technology may generate a considerable proportion of false positives among matches. This is known as a 'false positive paradox'.[2] Where a false positive paradox arises, providers may need to consider the increased likelihood that a positive detection could be a false positive.

1.20 The appropriate level of human review depends not only on the performance of the hash matching technology but also on the effectiveness of other technical and human review layers. While further review may be warranted where content is detected, it is not necessarily appropriate to prioritise human review solely based on a higher likelihood of false positives. Providers should instead consider how human judgement can meaningfully add value beyond automated systems and prioritise review where it is most likely to affect content moderation outcomes. For example, a provider might use more than one perceptual hash-matching algorithm (reducing the likelihood that each algorithm results in a false positive for a particular item of content considerably); or use machine-learning classifiers to identify items of detected content that are more or less likely to be false positives. Where the provider has a lower level of assurance in the detection outcomes

[2] Hunt, T., Williams, B. and Howard, D. 2021, The "false positive paradox" and the risks of testing asymptomatic people for COVID-19, Authorea, 25 June.

produced by the hash matching technology, we expect a higher proportion of matches to be subject to human review.

1.21 CSAM is a severe harm, both to those depicted in the content and those who may encounter it. Human review of content usually takes more time than automated processes do. Therefore, where human review is required, it may be appropriate for the service provider to temporarily remove the content from the service so that harm is not caused while it awaits human review. The content may be reinstated if, on further review, it is determined not to amount to CSAM.

1.22 The consequences for a user found to have posted CSAM can be severe. They are likely to include permanent blocking from the service concerned (see Code measure ICU H3) and reporting of the content to the NCA or relevant foreign agency. Providers should take particular care to minimise the reporting of false positives to the NCA or relevant foreign agency.

### Intimate image abuse hash matching

1.23 To determine what proportion of detected intimate image abuse content is appropriate for human review, providers should take into account:

a) Their level of assurance in the detection outcomes produced by the hash matching technology. For example, where perceptual hash matching is configured to prioritise recall, and/or where unverified hashes are used, a greater proportion of human review is likely to be appropriate to mitigate the increased risk of false positives, unless other technologies are used to verify the content;
b) The potential severity of the harm to those depicted in or encountering intimate image abuse content. For example, where human review may lead to a delay in content take down and the provider has a high level of assurance in the detection outcomes, it may be appropriate to have a lower proportion of human review;
c) The likely impacts on the user who generated, shared or uploaded the detected content or on the interested person (where applicable), if the proactive technology is incorrect, for example:

i) Restrictions on a user's access to the service, or their ability to use the service;
ii) An adverse impact on the user or interested person's financial remuneration payable in connection with, or generated by, the detected content (such as advertising revenue, subscriptions, commission or other monetisation arrangements), where known;
iii) the degree of any deprioritisation of the content in the overall ranking of the search results for UK users.

## Human review for quality assurance purposes

1.24 Human review of detected content for quality assurance may involve reassessing a sample of detected content and content moderation decisions relating to such content (including both automated and human decisions). The purpose of this type of human review is not to determine individual outcomes of moderation, complaints or appeals in real time, but to evaluate how the hash matching technology and related systems perform in practice, identify any systemic issues and correct any identified errors.

1.25 More human review for the purposes of quality assurance is likely to be needed where the level of assurance the provider has in the detection outcomes produced by the technology and any associated systems and processes – including other layers of proactive technology

and human moderators – is lower. Providers should also increase quality assurance, including human review as a part of quality assurance, where observed outcomes are inconsistent with expectations.

1.26 A proper approach to the configuration and ongoing quality assurance of hash matching technology is also likely to require human review of undetected content. This may include content which was not detected by the hash matching technology but has been identified through user reporting, complaints, community moderation, stratified sampling or other means. Where such content is subsequently assessed as illegal content that the hash matching technology was intended to detect under its current configuration, this constitutes a false negative.

1.27 When deciding the appropriate proportion of detected content for human review as part of quality assurance, providers should have regard to the potential impact on users or depicted persons arising from a failure to detect illegal content, including the risk that such content remains available to be encountered by UK users.