Draft Guidance on appropriate proportion of human review for intimate image abuse hash matching

Statement

Published: 18 May 2026 [To be finalised when issuing Code amendments]

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

About this Guidance

What this guidance covers

This is our guidance to assist providers of regulated user-to-user and search services in considering the appropriate proportion of content subject to human review when implementing hash matching processes for the purposes of detecting intimate image abuse.

The measures

1.1 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.

1.2 These measures say that such providers should ensure human moderators review and assess an appropriate proportion of detected content.

1.3 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.

1.4 For content and search moderation purposes, they should also have regard to:

a) the potential severity of the harm to those depicted in or encountering intimate image abuse content; 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).

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

1.6 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

1.7 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.

What should providers do?

1.8 For this measure, there are two purposes for human review:

a) Moderation: service providers may need to use human review to determine 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); and
b) Quality assurance: using human moderators for quality assurance and monitoring of automated technology.

1.9 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 content and search moderation purposes

1.10 To determine what proportion of detected 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. The level of assurance reflects the degree of confidence that a provider can reasonably place in the system as a whole;
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, 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.

1.11 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.

Human review for quality assurance purposes

1.12 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.

1.13 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.

1.14 Providers should also increase quality assurance, including human review as a part of quality assurance, where observed outcomes are inconsistent with expectations.

1.15 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.

1.16 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.