The Safety Net – Illegal Harms

Have you shielded your service?

As the OSA made its way through Parliament, the provisions aimed at tackling online 'illegal harms' sparked considerable discussion. This discussion focused on the material the government proposed to classify as 'harmful' and whether this definition was too broad, and also on whether the far-reaching nature of the content removal obligations could result in censorship of online content (potentially stifling freedom of expression and information).

There is no denying that the illegal harms requirements set out in the OSA are extensive and onerous and that there is a lot for in-scope service providers to get to grips with.

In line with its obligations under the OSA, Ofcom, the regulator for online safety in the UK, has published helpful material which, while detailed, explains the practical steps that Ofcom expects in-scope service providers to adopt in order to comply with these 'illegal harms' requirements. These materials aim to provide some clarity in terms of the compliance measures service providers are expected to implement, including how they should approach: (1) categorising content as illegal, (2) assessing the risks of that content, and (3) determining mitigating measures to adopt to limit the impact of the risks. The materials include:

- Regulatory Documents and Guidance which summarise and clarify the measures that in-scope service providers should implement to tackle illegal harms covering matters such as risk assessments, illegal content judgments, and enforcement.
- Codes of Practice on illegal content for user-to-user ('U2U') and search services which set out the steps that in-scope service providers should take to prevent users from encountering illegal content. While adherence to the codes is not mandatory, following them offers a 'safe harbour' for compliance with the OSA's requirements, meaning that compliance with the relevant OSA provisions is essentially guaranteed. Service providers may choose alternative compliance measures, but they must ensure these are equally effective in mitigating illegal content risk, as they will not benefit from the safe harbour.

In this article, we aim to assist in-scope service providers in digesting the available material and understanding their obligations under the OSA, by separating them into three distinct phases.

Importantly, the guidance and codes of practice were published in final form on 16 December 2024 and in-scope services have until **16 March 2025** to complete their illegal content risk assessments, following which they will need to have implemented all necessary measures to comply with the illegal harms requirements under the OSA (see Phase Two below for further information).

In summary, the Illegal Content Codes of Practice offer specific, actionable recommendations for compliance, while the Illegal Content Guidance provides overarching principles and clarifications to assist platforms in effectively implementing the OSA's provisions.

Phase One: Defining 'illegal content'

A fundamental aspect of 'illegal harms' compliance for in-scope services will be to understand what 'illegal content' is and how to identify it. Essentially, this is content which 'amounts to a relevant offence'; however, determining whether something constitutes a 'relevant offence' is not a straightforward assessment. The OSA therefore permits in-scope services to make 'illegal content judgements' if there are 'reasonable grounds to infer' that the content amounts to a relevant offence. This threshold of 'reasonable grounds to

infer' is lower than the threshold ('beyond reasonable doubt') that is applied in the UK's criminal justice system and may therefore result in a broader range of content being subject to moderation requirements.

This illegal content judgement should be made using all 'relevant information that is reasonably available'. There is no express requirement for in-scope services to carry out these judgements; however, they can be a useful way of understanding whether any additional OSA 'illegal harms' obligations are likely to apply to the service.

To assist with carrying out this judgement, Ofcom has published guidance on 'judgement for illegal content' which explains how providers can judge whether content is likely to be illegal. In this guidance, Ofcom lists 'priority' and 'non-priority' offences which would *all* constitute illegal content, but which carry with them different content moderation duties (see Phase Two for further detail). Schedules 5, 6, and 7 of the guidance list over 130 examples, which we have not listed; however, the offences can be broadly categorised as set out below:

Priority Offences	**Non Priority or 'Other' Offences**
• Terrorism • Harassment, stalking, and threat of abuse • Coercive and controlling behaviour • Hate offences • Intimate image abuse • Extreme pornography • Child sexual exploitation and abuse (CSEA) • Sexual exploitation of adults • Unlawful immigration • Human trafficking • Fraud and financial offences • Proceeds of crime • Assisting or encouraging suicide • Drugs and psychoactive substances • Weapons offences • Foreign interference • Animal welfare	• Epilepsy trolling (sending or showing flashing images) • Cyberflashing (sending photographs of genitals) • Encouraging or assisting serious self-harm • False communications • Threatening communications • Improper use of public electronic communications network

The above are categories of offences only and the guidance provides further key indicators and examples of each offence. There is an express obligation in the OSA requiring that the terms and conditions of the service provider's website should prohibit all of the above categories of content and a recommendation that providers may also wish to prohibit other categories of content.

Phase Two: Assessing the risks of illegal content

The OSA imposes obligations on *all* in-scope services (whether U2U or search) to carry out illegal content risk assessments. The purpose of the risk assessment is to improve a service provider's understanding of how risks of different kinds of illegal harm (whether physical or psychological) could arise on the service, and what safety measures need to be put in place to protect users. These obligations vary depending on whether the service is search or U2U. In its guidance on risk assessments, Ofcom recommends adopting the below four step methodology to illegal content risk assessments.

- **Step One:** Understand the harms that need to be assessed. As part of this step, in-scope service providers should consider the 17 different types of priority illegal content (see Phase One above) and identify whether there is a risk of each taking place on the service. In-scope services should consult Ofcom's risk profiles (see page 32 of the guidance) for each priority offence. Further, U2U services should understand how a service may be used to facilitate or to commit a priority offence.

- **Step Two:** Assess risks by considering the likelihood and potential impact of harms occurring on the service. This step will require providers of in-scope services to assess the likelihood and impact of each of the 17 priority offences, assess the different ways in which the service is used, identify any service-specific characteristics which may increase the risk, and consider the effectiveness of any existing controls that have been implemented. A 'risk level' should also be assigned to each of the 17 priority offences (see page 58 of the guidance). The risk level should be based on evidence inputs relevant to the service (e.g. user data including age, details of illegal content being complained about, results of product testing, and results of research / views of experts).

- **Step Three:** Implement safety measures and record outcomes of the risk assessment. As part of this step, in-scope service providers should consult Ofcom's codes of practice (see here for U2U and here for search) and determine which measures are proportionate and recommended for the relevant service and whether any additional measures may be appropriate (see Phase Three below for further details on this exercise). If providers can demonstrate that they comply with the codes, they will benefit from a 'safe harbour' and compliance with the OSA will essentially be guaranteed. In-scope service providers can choose to comply with these requirements using alternative measures to those outlined in the codes (provided of course these are equally effective in mitigating illegal content risk).

- **Step Four:** Report, review and update the risk assessment, at least every 12 months. A core aspect of OSA compliance is to ensure that all risk assessments are continuously reviewed, updated, and in particular that the effectiveness of the safety measures is monitored.

In each of the four steps listed above, service providers should also ensure that accurate written records are maintained, which not only evidence the conclusion reached, but also the various service characteristics that were taken into account (see Ofcom's specific record keeping and review guidance). Risk assessments should be retained in line with a service provider's document retention policy or for three years, whichever is longer. In some cases, depending on changing circumstances or amendments to Ofcom's 'Risk Profiles' in its guidance on risk assessments, it may be necessary to carry out new illegal content risk assessments.

Phase Three: Mitigating against the risk of illegal harms - key OSA duties

A core aspect of OSA compliance will be to implement appropriate measures to address any risks identified in the illegal content risk assessment. These measures are listed in Ofcom's codes of practice which should be consulted by in-scope service providers (there are separate codes for search and U2U services). At a minimum, all in-scope services (whether U2U or search) should: (1) have strong content moderation practices; (2) better facilitate user empowerment through easy to operate complaints processes; and (3) implement a governance framework to oversee and ensure the effective implementation of risk mitigation measures. Ofcom recommends the following broad categories of mitigating measures:

- **Search & Content Moderation:** Both search and U2U service providers should implement general content moderation measures to swiftly remove, index, and re-rank illegal content.
  Further, *larger* providers of medium and high risk services will be required to make use of automated tools to make content moderation processes more effective and efficient. Ofcom expects that using

such tools may require a tailored approach depending on the organisation and the type of content at issue, as set out in the below table.

Type of content	**Ofcom's recommendations**
CSAM	Proactive detection tools e.g., hash-matching and URL tracking.
Hate speech and extremism	Carefully consider context to distinguish between legal expressions and illegal incitements.
Fraud and financial crime	Dedicated reporting channels for trusted flaggers (e.g. government agencies, regulators). Keyword tracking was suggested as a possible solution in the draft guidance, but following consultation responses that this was too rudimentary a tool, Ofcom is considering other appropriate measures to mitigate this risk.
Terrorism	Monitoring known symbols, language patterns, and accounts linked to proscribed organisations. Ofcom may also recommend hash-matching in future.

- **User empowerment:** All in-scope services must have a complaints process and must take appropriate action in response to a UK user complaint. The process should be easy to use and must ensure that the complaint is acknowledged and that actions are taken in response to a complaint (e.g. removal or reinstatement of content, and an appeals process). Default privacy settings should be pre-emptively imposed on any child users of in-scope services. On the question of whether users who post illegal content should be blocked from accessing U2U services (rather than their content solely being removed), Ofcom recommends that an account should only be blocked where there are reasonable grounds to infer it is operated on behalf of terrorists. Finally, large and high risk services should ensure user controls are available to enable muting and comment-disabling, where appropriate.
- **Service Design:** Ofcom has provided the following substantive recommendations about the *design* of large horizontal search services (platforms that index and provide access to a wide variety of online content), when considering illegal harms:
    - predictive search functionalities (where used) should offer the ability to easily report predictive search suggestions which appear to direct users towards priority illegal content;
    - crisis prevention information should be provided by platforms in response to search requests that contain queries regarding suicide and suicide methods; and
    - providers should have different means to detect and warn against search requests relating to CSAM.
- **Terms of Service:** All in-scope service providers should have clearly signposted, easy-to-access, and comprehensible terms of service which:
    - explain how individuals will be protected from illegal content;
    - provide information about proactive technology used to detect and moderate illegal content; and

- explain the policies and processes which govern the handling of complaints regarding the presence of illegal content.

Additionally, providers of 'Category 1' services (see our article on the scope of the act) for more information regarding categorised services should summarise the findings of their illegal content risk assessment in the terms of service.

- **Governance & Accountability:** *All* in-scope services will need to ensure that a governance framework is in place to ensure that risk management activities are reviewed on an annual basis. Further, all in-scope services should name an individual accountable to the most senior governance body for compliance with the illegal content safety duties and the reporting and complaints duties. For large and multi-risk services only (see our article on the scope of the act) for more information regarding these service types), additional governance and accountability measures should be adopted such as:
    - tracking evidence of new and increased harms;
    - putting in place a code of conduct that sets standards and expectations for employees around protecting users from online harm; and
    - ensuring that relevant personnel involved in the design and operational management of the service are trained in the service provider's approach to compliance with the illegal content safety duties.

Conclusion & Next Steps

As the first duties go live and become enforceable, Ofcom has confirmed it will be proactively driving compliance with the rules in the following ways:

- through supervisory engagement with the largest and riskiest providers to ensure they promptly implement compliance measures;
- pushing for improvements where needed;
- gathering and analysing the risk assessments of the largest and riskiest providers;
- monitoring compliance and taking enforcement action if providers fail to complete their illegal harms risk assessment by 16 March 2025;
- early, focused engagement with certain high-risk providers to ensure they are complying with Ofcom's CSAM hash-matching measure, which recommends the use of effective hash-matching measures to detect and remove CSAM; and
- further targeted enforcement action for breaches of the safety duties where serious ongoing issues are identified that represent significant risks to users.

Ofcom has also confirmed that it will take steps to strengthen the codes with an additional consultation to take account of further measures currently being explored (e.g., AI for detecting illegal content and hash-matching measures).

For now, in-scope services should prioritise ensuring that the relevant illegal content risk assessments are completed before the 16 March 2025 deadline - to this end, the final guidance and codes of practice recently published by Ofcom will be an invaluable resource. These assessments are not merely a paper exercise and

in-scope services should also allow time for the corresponding mitigating measures identified as part of these assessments (and recommended by the codes of practice) to be tried, tested, and effectively implemented.

On 3 March 2025, Ofcom launched a new enforcement programme to monitor compliance with the OSA illegal content risk assessment duties. As part of this, Ofcom asked a number of large services, and some higher risk smaller services, to submit copies of their risk assessments by 31 March 2025, making it clear that any failure to do so could result in enforcement action.

Last updated: 5 March 2025