About this guidance

In detail

Why have you produced this guidance?

This guidance explains how data protection law applies when you use content moderation technologies and processes. It provides practical advice to help you comply with the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018 (DPA 2018). Read it to understand the law and our recommendations for good practice.

It is not a comprehensive guide to compliance. We link to relevant further reading about any principles we’ve already covered in our other guidance.

If you are processing children's personal information, you **should** conform with our Children's code. When we refer to a child we mean anyone under the age of 18. Our code is a statutory code of practice that sets out how internet society services likely to be accessed by children can protect children’s information rights online. It sets out fifteen standards that you **should** implement, if you are an internet society service.

The Data Protection and Digital Information Bill was reintroduced in the Houses of Parliament on 8 March 2023. When the Bill becomes law, it will amend elements of the DPA 2018 and the UK GDPR relevant to this guidance. We have written this guidance in line with the applicable law at the time of publication.

This guidance on content moderation is the first in a series of products we’ve planned about online safety technologies.

This is part of our ongoing commitment to publishing guidance on online safety technologies, alongside our work to ensure regulatory coherence between the data protection and online safety regimes. We announced this in our 2022 joint statement with Ofcom on online safety and data protection.

Who’s it for?

This guidance is for organisations that use or are considering using content moderation. It is also for organisations that provide content moderation products and services. It is for both data controllers and processors.

The guidance is aimed at organisations who are carrying out content moderation to meet their obligations under the Online Safety Act 2023 (OSA). However, it is also applies to organisations who are carrying out content moderation for other reasons.

Whether you are carrying out content moderation to comply with the OSA or for other purposes, you **must** comply with data protection law.

We expect that this guidance will be most relevant to trust and safety professionals. It will also be relevant to those in roles with a data protection compliance focus, such as data protection officers, general counsel, privacy-legal professionals and risk managers.

What does it cover?

It sets out how organisations deploying content moderation processes or providing content moderation services can comply with data protection law.

In this guidance we define 'content moderation' as:

- the analysis of user-generated content to assess whether it meets certain standards; and
- any action a service takes as a result of this analysis. (See the section 'What do we mean by content moderation?' for more information.)

This guidance focuses on moderation of user-generated content on user-to-user services. For the purposes of this guidance, we follow the definitions in the OSA.

Section 3(1) of the OSA defines a user-to-user service as:

> "User-to-user service means an internet service by means of which content that is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, may be encountered by another user, or other users, of the service."

Section 55(3) of the OSA defines user-generated content as:

> "User-generated content means content that is:
>
> - (i) generated directly on the service by a user of the service, or (ii) uploaded to or shared on the service by a user of the service, and;
> - that may be encountered by another user, or other users, of the service by means of the service."

This guidance discusses content moderation processes that are managed and administered by organisations. It applies to content moderation that is manual, partly automated and solely automated.

What doesn't it cover?

This guidance sets out the requirements of data protection law where you process personal information in content moderation. It does not cover the following:

- Compliance with the specific obligations in the OSA. The regulator for the online safety regime is Ofcom. Please consult Ofcom's codes of practice and guidance for information about what you are required to do under the OSA.
- The use of behaviour identification and user profiling. In many cases, you may use content moderation alongside other systems and processes. This includes those that analyse the behaviour of users on a service, or build a profile of them to assess their characteristics. We plan to produce guidance on this in the future.
- Specific considerations that arise from on-device moderation, such as the application of the Privacy and Electronic Communications Regulations (PECR).
- The requirement in section 66 of the OSA for regulated services to report all detected and unreported Child Sexual Exploitation and Abuse (CSEA) content to the National Crime Agency (NCA). We will publish further data protection guidance about this when these regulations are implemented.

How do we use this guidance?

To help you to understand the law and good practice as clearly as possible, this guidance says what organisations **must, should,** and **could** do to comply.

Legislative requirements

- **Must** refers to legislative requirements.

Good practice

- **Should** does not refer to a legislative requirement, but what we expect you to do to comply effectively with the law. You should do this unless there is a good reason not to. If you choose to take a different approach, you must be able to demonstrate that this approach also complies with the law.
- **Could** refers to an option or example that you could consider to help you to comply effectively. There are likely to be various other ways you could comply.

This approach only applies where indicated in our guidance. We will update other guidance in due course.

We plan to keep this guidance under review and update it where appropriate, for example to reflect Ofcom’s final online safety codes of practice and guidance.

How does this guidance relate to the OSA?

The OSA sets out rules for user-to-user and search services. These services have new duties to protect UK users by assessing and responding to risks of harm. This includes duties on user-to-user service providers to:

- use proportionate measures to prevent users from encountering certain types of illegal content; and
- use proportionate systems and processes to swiftly remove any illegal content after becoming aware of

its presence on the service.

If a service is likely to be accessed by children, the OSA sets out duties for the protection of children. The OSA also includes specific duties for services that display or publish provider pornographic content.

Ofcom is the regulator for the OSA. It is responsible for implementing the regime and supervising and enforcing the online safety duties. Ofcom is publishing codes of practice and guidance that will provide more detail about the regime and explain how you can comply with your new duties.

The OSA sits alongside data protection law. Compliance with one does not necessarily mean compliance with the other. If you are carrying out content moderation that involves personal information, you **must** comply with data protection law.

Further reading

- Children's code including the section on Services covered by this code.
- Ofcom OSA codes and practice and guidance.
- Online Safety Act 2023.

What is content moderation and how does it use personal information?

In detail

What do you mean by content moderation?

We use the term 'content moderation' to describe:

- the analysis of user-generated content to assess whether it meets certain standards; and
- any action a service takes as a result of this analysis. For example, removing the content or banning a user from accessing the service.

You may carry out content moderation for a range of purposes, including meeting your obligations under the OSA or enforcing your terms of service.

In this guidance, we use the term 'content policies' to describe the rules you set out in your terms of service that specify:

- what content you do not allow on your service; and
- how you deal with certain types of content.

Content policies usually prohibit content that is illegal, as well as content that you deem to be harmful or undesirable.

When we discuss 'moderation action' in this guidance, we are referring to action you take on a piece of content or a user's account after you've analysed the content. This may be because you are:

- required to take action to comply with your duties under the OSA; or
- enforcing your content policies.

For example:

- **Content removal** – you may remove content from your service (or prevent it from being published, if moderation is taking place pre-publication).
- **Service bans** – you may ban users from accessing your service, either temporarily or permanently. You

may operate a ‘strike’ system that records content policy violations by a user, and enforces a ban when a user reaches a certain number of strikes.

- **Feature blocking** – you may restrict a user’s access to certain features of your service, either temporarily or permanently. For example, you may block users from posting content, or from commenting on content posted by others, while still giving them access to other features of the service normally.
- **Visibility reduction** – a range of actions you may take to reduce the visibility of content. For example, you may prevent content from being recommended or make content appear less prominently in users’ news feeds.

What are the key stages in content moderation?

These are the key stages a content moderation workflow may involve:

- **Database matching** – an automated analysis of content to check whether it matches an internal or external database of known prohibited content. For example, hash matching is a type of database matching technology that is commonly used for detecting exact or close matches of known child sexual abuse material (CSAM).
- **Content classification** – an automated analysis of content to assess whether it is likely to breach a service’s content policies. This often uses artificial intelligence (AI) based technologies. Classification systems may assign a degree of confidence to their assessment of a piece of content. Services may decide that if content reaches a particular certainty threshold, it can be automatically actioned (eg removed or queued for human review).
- **Human review** – human moderators reviewing content against a service’s content policies. This includes content that has been flagged by an automated tool and content that has been reported by other people, such as other users or third parties.
- **Moderation action** – taking action on a piece of content (eg removing it from the service) or a user’s account (eg banning the user from the service). (See the section on ‘What do you mean by content moderation?’ for more information.)
- **Appeals and restorations** – the processes that allow users to challenge moderation decisions. Typically this involves human review of the content and the moderation decision.

However, there are different approaches to content moderation and the stages may differ. This means you may have configured your systems to:

- only follow some of these stages. For example, you may not human review all content before deciding to take moderation action;
- rely on users or third parties to report content for human review, rather than using database matching or content classification;
- take moderation action before or after you publish the content; or
- make decisions automatically using content moderation technology, or by a human moderator, or both. (See the section on ‘What if we use automated decision-making in our content moderation?’ for more information.)

Sometimes you may be able to make decisions about content based solely on the content itself. However, you may also need to look at other personal information associated with a user’s account. (See the section below on ‘What personal information does content moderation involve?’ for more details.)

How might third parties be involved in content moderation?

You may use third-party content moderation providers to assist you with a range of processes:

- **Third-party technology** – they can provide you with a range of automated moderation technologies. They may have expertise in moderating specific formats or categories of content.
- **Third-party human moderation** – they can provide human moderators to review content that has been flagged as potentially violating your content policies. They may also manage other processes, such as user appeals.

If you use third-party providers, you **must** make sure that all parties clearly understand whether they are acting as a data controller, joint controller, or processor. (See the section on 'Who is the controller in our content moderation systems?' for more information.)

What personal information does content moderation involve?

Personal information means information that relates to an identified or identifiable person.

This doesn't just mean someone's name. If you can distinguish a person from other people, then they are "identified" or "identifiable". Examples of information that may identify someone include an IP address or an online identifier.

Content moderation systems involve processing people's personal information at all stages of the workflow.

In most cases, user-generated content is likely to be personal information in your moderation systems. This can be because:

- it's obviously about someone. For example, if the content contains information that is clearly about a particular user of your service; or
- it's connected to other information, making someone identifiable. For example, the account profile of the user who uploaded it, which may include information such as their name, online username and registration information (eg email address).

Content moderation may also involve using personal information that is linked to the content or a user's account. For example, a user's age, location, previous activity on the service, or a profile of their interests and interactions.

You **must** be able to demonstrate that using this kind of personal information in content moderation:

- is necessary and proportionate; and
- complies with the data minimisation principle. (See the section on 'How do we ensure data minimisation in our content moderation?' for more information).

Further reading

- What is personal information: a guide

What do we do if we use pseudonymised personal information in our content moderation?

In some content moderation systems, you may be able to separate the content from information you hold about the user who uploaded it. You may do this to analyse the content at the detection and classification stages.

The information you analyse is not truly anonymised in these circumstances. Instead, you are processing pseudonymised personal information.

Pseudonymisation has a specific meaning in data protection law. This may differ from how the term is used in other circumstances, industries or sectors.

The UK GDPR defines pseudonymisation as:

> "…processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person."

Pseudonymisation therefore refers to techniques that replace, remove or transform information that identifies a person. For example, replacing one or more identifiers which are easily attributed to a person (eg names) with a pseudonym (eg a reference number or job ID).

While you can link that pseudonym back to the person if you have access to the additional information, you **must** hold this information separately. You **must** have technical and organisational measures in place to ensure you can do so.

Although you cannot directly identify people from pseudonymous information, you can identify them by referring to other information you hold separately. Therefore any information you have pseudonymised remains personal information and you **must** comply with data protection law when processing it.

Example

An online video gaming service hosts a forum where users can discuss video games and share pictures and videos. The content uploaded by users includes discussion about their experience playing games and images and videos of their gameplay.

The service employs a team of moderators to ensure that the forum posts don't violate its content policies. This involves the moderation team processing the forum users' personal information.

The service has another team that is responsible for scanning the content that users upload to the forum. This team doesn't need to identify individual users, as its purpose relates to the content specifically.

The service ensures that this second team doesn’t receive information that can identify individual users. It pseudonymises this information by replacing the users’ names and account handles with a reference number. This helps to reduce risks and improve security, while still enabling the service to fulfil this particular purpose.

The service is still processing personal information in both cases, even if it applies technical and organisational measures to ensure the second team can’t identify people without the additional information. This is because the organisation can, as the controller, link that material back to the identified users.

Further reading

- Draft guidance on anonymisation and pseudonymisation
- Privacy-enhancing technologies (PETs)

Do content moderation systems use special category information?

Content moderation systems may process special category information about people. This means personal information about a person’s:

- race;
- ethnic origin;
- political opinions;
- religious or philosophical beliefs;
- trade union membership;
- genetic data;
- biometric data (where this is used for identification purposes);
- health data;
- sex life; or
- sexual orientation.

The UK GDPR is clear that special category information includes not only personal information that specifies relevant details, but also personal information ‘revealing’ or ‘concerning’ these details.

Content moderation may involve processing special category information both directly and by inference. This could be if:

- you use special category information about users to support the decisions you make about their content (eg because it provides additional context);
- you are intentionally inferring details that fall within the special categories of information to inform the outcome of your content moderation (see the example in the box below); or

- the content you are moderating includes special category information about users. For example, posts on an online forum where identifiable users are directly expressing their political views. If you are moderating user-generated content that contains this information, then you are processing special category information regardless of whether you intend to.

You **must** ask yourself whether you are likely to use any special category information (including inferences) to influence or support your activities in any way. If so, then you are processing special category information. Special category information needs more protection because it is sensitive. You **must** identify a condition for processing, in addition to a lawful basis, if you are processing it, either because you've planned to or because it's contained within the content. (See the section on 'How do we carry out content moderation lawfully?' for more information.)

In some cases, you may be able to surmise details about someone that fall in to the special categories of information, even though you do not intend to make those inferences. For example, a human moderator reviewing images and videos of people wearing certain clothing may be able to infer that they belong to a particular religious group, even if the content does not specify that information directly.

You are not processing special category information, if you do not:

- process the content with the purpose of inferring special category information; nor
- intend to treat people differently on the basis of an inference linked to one of the special categories of information.

However, as above, you are processing special category information if you intentionally use those inferences to inform your moderation. This is the case regardless of whether that inference is correct.

Example

A service deploys a content moderation system that analyses user-generated pictures and videos when they are uploaded, in order to detect content that is promoting self-harm or suicide.

If the analysis finds the content is likely to be promoting suicide or self-harm, it is referred to a human moderator for review. In some cases, where the moderator deems the user to be in immediate danger, the service shares the user's information with the emergency services.

This system is processing special category information (ie health related information indicating whether the person is at risk of self-harm or suicide). It is making a special category inference at the analysis stage, and sharing that special category personal information with the emergency services.

The service therefore needs to identify a lawful basis and a valid condition for processing, at both the analysis and referral stages.

Further reading

- Special category data

Is criminal offence information a relevant consideration?

The UK GDPR gives extra protection to personal information relating to criminal convictions and offences or related security measures.

This includes personal information ‘relating to’ criminal convictions and offences. For example, it can cover suspicion or allegations of criminal activity. In this guidance, we refer to this information collectively as ‘criminal offence information’, although this is not a term used in the UK GDPR.

You **must** assess whether you are processing criminal offence information as part of your content moderation.

Section 192 of the OSA requires services to take down content where they judge there to be “reasonable grounds to infer” it is illegal, using “reasonably available” information to make this judgment. Ofcom has produced draft guidance on how services can make illegal content judgements for the purposes of the takedown duty, the risk assessment duty and the safety duties more generally.

It depends on the specific circumstances of your processing as to whether you are processing criminal offence information about a person when you make an illegal content judgment under the OSA.

If you are carrying out content moderation that involves processing criminal offence information, you **must** identify a condition for processing, as well as your lawful basis. (See the section on ‘How do we carry out content moderation lawfully?’ for more information.)

We plan to publish further data protection guidance on reporting CSEA content to the NCA under section 66 of the OSA.

Further reading

- Criminal offence data
- Online Safety Act 2023

How do we assess and mitigate data processing risks?

Content moderation is a type of processing that is likely to result in a high risk to people's rights and freedoms. This means you **must** carry out a data protection impact assessment (DPIA) prior to the processing because your content moderation is likely to involve:

- processing involving new technologies, or the novel application of existing technologies (including AI);
- combining, comparing or matching personal information obtained from multiple sources;
- solely automated processing that has a legal or similarly significant effect on the user; or
- decisions about a person's access to a service based on automated decision-making or use of special category information.

You **must** carry out a DPIA if you are using children's personal information as part of offering an online service directly to them.

You **must** include the following in your DPIA:

- describe the nature, scope, context and purposes of the processing. Be clear about what personal information you want to process and why;
- assess necessity, proportionality and compliance measures; and
- identify all relevant risks to people's rights and freedoms, assess their likelihood and severity and detail measures to mitigate them.

You **should** also consider the types of data protection harms that these risks may lead to. For example, content moderation has the potential to lead to:

- loss of control of personal information through unexpected and unfair use or sharing of information;
- adverse effects on rights and freedoms, including privacy rights and rights to freedom of expression;
- financial harm to people (eg through loss of income or employment); and
- discrimination based on a moderation system's outputs.

You **should** carry out a DPIA, even if you assess your processing is not likely to result in high risk. This is because it is a flexible and scalable tool which can assist your decision-making and risk mitigation. If you decide to proceed without carrying out a DPIA, you **should** document your decision.

If you have carried out a DPIA that identifies a high risk that you cannot reduce to an acceptable level, you **must** consult us before going ahead with the planned processing.

You **must** follow a data protection by design and default approach when you decide to use a content moderation system. This helps you consider privacy and data protection issues at the design stage of your system, and throughout its operation. Following a data protection by design approach means you **must**:

- put in place appropriate technical and organisational measures designed to implement the data protection principles effectively; and
- integrate safeguards into your processing so you meet the UK GDPR's requirements and protect people's rights.

If you use automated decision-making in your content moderation systems, then there are data protection requirements that you **must** comply with. (See the section on ‘What if we use automated decision-making in our content moderation?’ for more information.)

Further reading

- Data protection impact assessments (DPIAs)
- Examples of processing likely to result in a high risk
- Children’s code – see standard 2 for carrying out a DPIA when processing children’s information.
- Overview of data protection harms and the ICO’s taxonomy
- Data protection by design and default
- Privacy in the product design lifecycle
- Guidance on AI and data protection

How do we carry out content moderation lawfully?

In detail

What does it mean for our content moderation processing to be lawful?

The first data protection principle requires any processing of personal information to be lawful, fair and transparent.

You **must** identify a lawful basis before you start using personal information in your content moderation system.

There are six available lawful bases for processing set out in Article 6 of the UK GDPR. No one basis is always better or more important than the others.

Your decision depends on the specific purposes you are going to use content moderation for.

In practice, the lawful bases that are most likely to be relevant to your content moderation processing are:

- **legal obligation** – you can rely on this lawful basis if you need to process someone's personal information to comply with a common law or statutory obligation. For example, you may be carrying out content moderation to comply with your safety duties under the OSA; and
- **legitimate interests** – you can rely on this lawful basis if you have a legitimate interest in processing someone's personal information. For example, you may be using content moderation to enforce your terms of service. This basis involves balancing your interests against the person's interests, rights and freedoms.

Although legal obligation and legitimate interests are the most likely lawful bases to apply in practice, we have included guidance if you are considering the remaining lawful bases.

If your content moderation system uses special category information or criminal offence information, you **must** have a lawful basis **and** an additional Article 9 or Article 10 condition for processing. (See sections below on 'What if our content moderation involves special category information?' and 'What if our content moderation involves criminal offence information?' for more information.)

Further reading

- A guide to lawful basis

What do we need to consider if we’re using legal obligation as our lawful basis?

You can use legal obligation as your lawful basis if you need to process personal information to comply with a common law or statutory obligation. The information in this section is about using legal obligation under the OSA. However, there may be other statutory obligations or common law duties that are relevant to you.

If you process personal information in your content moderation to comply with your obligations under the OSA, you are likely to be able to rely on legal obligation as your basis for this processing. But, the processing **must** be necessary and proportionate to achieve compliance.

You are also likely to be able to use this lawful basis for personal information processing that you need to do to apply the measures recommended in Ofcom's codes of practice under the OSA. This is because the codes provide measures that enable you to comply with the legal obligations set out in the OSA.

You **must** directly link your processing to a legal obligation placed on you. Your processing does not need to be essential for you to comply with your obligation, but you **must** ensure it is a reasonable and proportionate way of achieving compliance. You **should** document your decision to rely on legal obligation and identify which part of the legislation specifies this.

You **must not** rely on this lawful basis for content moderation processing that goes beyond what is required for you to meet your duties in the OSA (unless there are other common law duties or legislative obligations that apply).

Example

An online service establishes a content moderation system that involves human moderators analysing user-generated content to assess whether it breaches the service’s content policies. They remove content that breaches the service’s content policies. This system processes users’ personal information.

The service implements this measure in accordance with the recommended measure included in Ofcom’s draft illegal content codes of practice for user-to-user services.

The service may rely on legal obligation as its lawful basis for this processing provided that this a recommended measure in Ofcom’s final code of practice. This is because the service is carrying out the processing as part of complying with its duties under the OSA.

Further reading

- Legal obligation
- Online Safety Act 2023

What do we need to consider if we're using legitimate interests as our lawful basis?

Legitimate interests is the most flexible of the lawful bases. It isn't focused on a particular purpose, which means you can potentially rely on it in a range of circumstances.

Legitimate interests is most likely to apply where you want to use personal information in ways that:

- people would reasonably expect; and
- don't have an unjustified adverse impact on people's rights and freedoms.

If there is an impact on people, legitimate interests may still be available, but you **must** show that there is a compelling benefit to the processing and the impact is justified.

You are likely to have a legitimate interest in detecting and taking action on content in accordance with your content policies. However, you **must** balance this interest against the interests, rights and freedoms of users. You **must** also make sure that your personal information processing is necessary to achieve your purpose. If you can reasonably achieve the same result in another less intrusive way, legitimate interests does not apply.

To ensure you've considered these issues, you **should** carry out a three-part test to:

- identify your legitimate interest;
- show that processing is necessary to achieve that interest; and
- consider whether people's interests, rights and freedoms override the legitimate interest you've identified.

It's likely that your use of personal information in content moderation will involve a level of intrusion into your users' privacy. This means you **must** demonstrate that you have a compelling justification for this (eg safeguarding your users).

You can consider legitimate interests for processing children's personal information, but you **must** take extra care to protect their interests.

What you tell people in your privacy information is one of the factors that affects whether they can reasonably expect the processing. You **should** be clear with users about what types of content are prohibited on your service and why, and how you action certain content. If you provide this information, users are more likely to expect the processing that you undertake to detect and remove this type of content.

Further reading

- Legitimate interests
- Children and the UK GDPR

Can we use contract as our lawful basis?

You can use contract as your lawful basis if using personal information is objectively necessary to a deliver a contractual service to a relevant person, or to users of your service in general.

You **must not** rely on contract if:

- there is a less intrusive way of processing personal information to provide the same service; or
- the processing is not objectively necessary for the performance of the contract.

If you are thinking about using contract, you **should** consider the following questions:

- Are you carrying out the processing to deliver a contractual service? For example, are you processing personal information to fulfil obligations that you have outlined in your terms of service?

If your answer is yes:

- Is the processing **necessary** for the performance of that contract? For example, is processing the information a targeted and proportionate step that is integral to delivering the contractual service?

In most cases, although you may be able to rely on contract for your contract moderation processing, it is likely that legal obligation or legitimate interests are more suitable.

If the contract is with a child under 18, you **must** consider whether they have the necessary competence to enter into a contract. If you have doubts, you may wish to consider an alternative basis, such as legitimate interests. Using legitimate interests as your lawful basis can help you demonstrate that you properly considered and protected the child’s rights and interests.

You are unlikely to be able to rely on contract for processing personal information for purposes such as ‘service improvement’ of your content moderation systems. In most cases, collecting personal information about how people engage with a service in order to develop new service functions is not objectively necessary to provide a contract. This is because you can deliver the service without this processing.

Further reading

- Contract
- Children and the UK GDPR

Can we use consent, vital interests or public task?

It is unlikely that these lawful bases can apply to your content moderation processing.

Consent is about giving people genuine choice and control over their information. Consent won’t apply as you're unlikely to be offering people a free choice about whether you process their information for content moderation.

Vital interests generally only applies in specific matters of life and death. It is unlikely to apply to your content moderation processing, particularly at scale.

Public task is unlikely to be relevant to the user-to-user services this guidance applies to.

Further reading

- Consent
- Vital interests
- Public task

What if our content moderation involves special category information?

In order to lawfully process special category information, you **must** identify a condition for processing, as well as a lawful basis.

There are 10 conditions for processing special category information outlined in Article 9 of the UK GDPR. Five of these require you to meet additional conditions and safeguards set out in schedule 1 of the DPA 2018. In many cases you also need an 'appropriate policy document' in place in order to meet a schedule 1 condition in the DPA 2018.

When choosing a condition for processing, being clear about your purpose for content moderation will help you identify the most appropriate condition.

If you plan to use special category information, or if you are moderating content that includes special category information about users, then you **must** identify a condition for processing. (See the section of this guidance on 'Do content moderation systems use special category information?' for more information.)

If you are not sure whether the user-generated content you intend to moderate contains special category information, you **should** identify a condition for processing to cover that possibility and minimise the privacy risks.

Below, we discuss some of the conditions for processing special category information that may be relevant in content moderation.

Substantial public interest

In order to rely on this condition, you **must** demonstrate that your processing has substantial public interest benefits.

In order to demonstrate this, you must **meet** one of 23 specific substantial public interest conditions. For almost all of these conditions, you **must** have an appropriate policy document in place. The conditions set out in part 2, schedule 1 DPA 2018 that may be relevant include:

- **preventing or detecting unlawful acts** – this condition is met if your use of personal information is necessary to prevent or detect an unlawful act;
- **safeguarding of children and individuals at risk** – this condition applies if your use of personal information is necessary to protect a child or at-risk person from neglect or harm, or to protect their wellbeing; and
- **regulatory requirements** – this condition applies if your use of personal information is necessary to comply with a regulatory requirement that involves establishing whether someone has committed an unlawful act or has been involved in dishonesty, malpractice or other seriously improper conduct.

All of these conditions also require you to demonstrate that your specific processing of special category information is "necessary for reasons of substantial public interest".

Further reading

- Special category data including the specific section on What are the substantial public interest conditions?

What if our content moderation involves criminal offence information?

You **must** ensure that you process any criminal offence information lawfully, fairly and transparently, and that you have an Article 6 lawful basis for processing.

In addition, Article 10 of the UK GDPR states that you **must** only process criminal offence information if this processing is:

- under the control of official authority; or
- authorised by domestic law. In the UK, this means you need to meet one of the conditions in schedule 1 of the DPA 2018.

You are unlikely to be processing under the control of an official authority when carrying out content moderation (see our guidance on criminal offence information for more details). Therefore, you **must** identify a specific condition for processing in schedule 1 of the DPA 2018, if your content moderation processing involves criminal offence information.

As with special category information, you may require an appropriate policy document depending on the condition you rely on.

The following schedule 1 conditions may be relevant for processing criminal offence information in content moderation systems:

- preventing or detecting unlawful acts (see above);
- safeguarding of children and individuals at risk (see above); and
- regulatory requirements (see above).

Further reading

- Criminal offence data

How do we make sure our use of personal information is fair?

The fairness principle in data protection law means you **must** only process personal information in ways that:

- people would reasonably expect; and
- could not have unjustified adverse effects on them.

Fairness in a data protection context is about how you go about the processing and the outcome of the processing (ie the impact it has on people).

When carrying out content moderation, you **must not** process people's personal information in a way that they might find unexpected, misleading or unforeseen. Fairness is closely linked to transparency. Part of ensuring your processing is fair is expaining to users how you use their information. (See the section on 'How should we tell people what we're doing?' for more information.)

You **must** ensure that your content moderation systems perform accurately and produce unbiased, consistent outputs. You are unlikely to be treating users fairly if you make inaccurate judgements or biased moderation decisions based on their personal information.

You **should** regularly review how you use personal information in your content moderation processes to minimise the risk of unfair outcomes for users. For example, you **could** provide guidance and training to your moderators on how to make consistent and fair moderation decisions based on personal information. You **could** also audit moderator decisions periodically to check that they use personal information consistently and fairly.

It is particularly relevant to consider fairness if you are using AI technologies to help you analyse content. This is because AI technologies can be susceptible to bias, which can result in discriminatory outcomes.

You **must** ensure that any technologies you use to process personal information are sufficiently statistically accurate and avoid discrimination. You **should** conduct regular checks of your systems for accuracy and bias.

Further reading

- Principle (a): Lawfulness, fairness and transparency
- Guidance on AI and data protection, including the specific sections on 'What do we need to know about accuracy and statistical accuracy?' and 'How do we ensure fairness in AI?'

How should we tell people what we’re doing?

Data protection law requires you to inform people that you are processing their personal information. You **must** tell people about:

- how you use their information;
- what decisions you make using their information; and
- how they can exercise their data protection rights.

Transparency in your use of people’s personal information is closely linked to fairness. Your processing is unlikely to be fair if you do not provide people with information about it.

Regardless of the type of content moderation you use, if your systems are processing personal information, then you **must** tell users:

- why you are using their personal information;
- what lawful basis you are using for processing;
- what types of personal information you are using;
- what decisions you are making with the information and how it impacts their use of your service;
- whether you keep personal information used or generated by your systems and for how long;
- whether you share their personal information with other organisations; and
- how they can exercise their data protection rights.

You **must** provide information to users in a way that is accessible and easy to understand. This is particularly important if your service is likely to be accessed by children.

If your content moderation involves solely automated decision-making based on personal information that has a legal or similarly significant effect on the people involved, then you **must** explain:

- that you use automated decision-making processes;
- what information you use;
- why you use it; and
- what the effects on them might be.

(See the section on ‘What if we use automated decision-making in our content moderation?’ for more information.)

Under the OSA, providers of regulated user-to-user services are required to include provisions in their terms of service giving information about any proactive technology they use for complying with the illegal content safety duties or the safety duties for the protection of children. This includes providing information about content identification technology.

As well as complying with the information requirements of the OSA, you **must** comply with your transparency obligations under data protection law. Please consult Ofcom guidance for more information about your obligations under the OSA.

Further reading

- Principle (a): Lawfulness, fairness and transparency
- Accountability framework – transparency
- Children’s code – see standard 4 for how to provide information in an accessible and easy to understand way
- Right to be informed
- Online Safety Act 2023

How do we define our content moderation purposes?

Purpose limitation means that you **must** only collect personal information for specified, explicit and legitimate purposes. You **must not** further process it in a way that's incompatible with those purposes.

This data protection principle is closely linked to the others, including data minimisation; and fairness, lawfulness and transparency.

You **must** be clear from the outset:

- why you're using personal information for content moderation (eg to comply with your obligations under the OSA, or to enforce your content policies); and
- what you intend to do with that information.

You **should** regularly review your processing, documentation and privacy information to check that your purposes have not evolved, beyond those you originally specified.

If your purposes change over time or you want to use personal information for a new purpose which you did not originally anticipate, you **must** ensure that:

- the new purpose is 'compatible' with the original purpose;
- you get the person's specific consent for the new purpose; or
- you can point to a clear legal provision requiring or allowing the new processing in the public interest.

You **must** have a lawful basis for your new purpose. This may be different from the original lawful basis you used to collect the information. (See the section on 'How do we carry out content moderation lawfully?' for more information.) You **must** update your privacy information to make sure that your processing is transparent.

In some cases, your third-party moderator(s) may want to use the personal information collected and generated during content moderation for other purposes. For example, developing and improving content moderation products.

You **should**:

- establish whether any moderation service provider you intend to use wants to do this;
- confirm that the provider would be acting as a controller for this particular use of personal information. (See section on 'Who is the controller in our content moderation systems?' for more information); and
- establish how you, or the provider, would inform people about the use of their personal information for these purposes by the third party.

You **should** regularly review any services you outsource and be able to modify or switch them to another provider if their use of personal information no longer complies.

Further reading

- Principle (b): Purpose limitation

How do we ensure data minimisation in our content moderation?

The data minimisation principle means you **must** use personal information in content moderation in a way that is adequate, relevant and limited to what is necessary.

You **must** take particular care when processing a child's personal information.

Content moderation technologies and methods are capable of gathering and using more information than may be necessary to achieve your purposes. This risks unnecessary intrusion into your users' privacy.

In many cases you can make accurate content moderation decisions based solely on the content. If so, you **must** avoid using other personal information associated with the content or user's account.

Moderation of content can be highly contextual. Sometimes, you may need to use other types of personal information (beyond just the content) to decide whether you need to take moderation action, including users':

- previous posts on the service;
- records of previous content policy violations;
- interactions on the service (such as likes and shares); and
- interests.

You are complying with the data minimisation principle, as long as you can demonstrate that using this information is:

- necessary to achieve your purpose (eg because it ensures your decisions are accurate and fair); and
- no less intrusive option is available to achieve this.

You **must** be clear about:

- what personal information you anticipate is necessary to make decisions about content on your service; and
- the circumstances when you might need to use this information.

You **should**:

- document this and be able to justify it;
- keep your record of this under review in case you determine further types of personal information as being necessary in future;
- consider using pseudonymisation to achieve data minimisation, where appropriate. Pseudonymisation helps reduce the risks to people and improve security;
- provide clear guidance and training for your moderators (including any volunteer or community moderators you may use). This will help them understand what personal information to use in their decision-making and the requirements of data protection law; and
- ensure that moderators understand when to escalate decisions about content.

You **could** implement access controls to ensure that human moderators are only able to view and access personal information that is relevant to inform moderation decision-making. For example, using an interface that only displays the content to a human moderator in the first instance, with the option to apply for access to additional user details, if required.

If you are using third-party moderation providers, you **must** limit the information you give them to what is relevant and necessary for them to carry out moderation.

Example

A service deploys a content moderation process that aims to detect content that violates its content policies.

A piece of content is flagged as potentially violating the services content policies and the system sends it to a human moderator for review.

Initially, the moderator reviews the content on an interface that displays only the content.

In this case, the content alone is not sufficient to make an accurate judgement about it. Therefore, the moderator needs to analyse additional personal information to support their decision.

The moderator consults guidelines provided by the service that explain what additional personal information may be necessary for their decision-making.

Following the guidelines, the moderator applies for access to view the user’s previous posts in the thread and their moderation history on the site. The moderator’s access to this information is logged by the service.

This represents good practice under the data minimisation principle. This is because the information used by the moderator is kept to the minimum needed for them to make an accurate decision.

Data minimisation and illegal content judgements under the OSA

Data minimisation still applies when services use personal information to make illegal content judgements under section 192 of the OSA. Under data protection law, this means you **must** use personal information that is proportionate and limited to what is necessary to make illegal content judgements.

To comply with data protection law, you **should**:

- be clear, in advance, about what personal information you may need to make illegal content judgments;
- document this and keep it under review; and
- provide clear guidance for your content moderators on using personal information to make illegal content judgements.

Further reading

- Principle (c): Data minimisation
- Children’s code – see standard 8 for more information on data minimisation
- Draft guidance on anonymisation and pseudonymisation
- Online Safety Act 2023

How do we ensure the accuracy of personal information?

The accuracy principle in data protection law means you **must**:

- take all reasonable steps to ensure the personal information you use and generate through your content moderation processes is not incorrect or misleading as to any matter of fact;
- keep the personal information up-to-date, if necessary; and
- consider any challenges from users about the accuracy of the information you've gathered through your content moderation.

For example, if you determine that a user has violated your content policies and you record this on their account, you **must**:

- ensure that this record is accurate;
- keep it up-to-date, where necessary; and
- take reasonable steps to rectify or erase the record without delay, if a user successfully appeals that they have not breached your content policies.

Further reading

- Principle (d): Accuracy

How long should we keep personal information for?

You **must not** keep personal information obtained from your content moderation activities for longer than you need it. There are no set time limits in data protection law because it depends on your situation and your purposes for processing the information. You **must not** hold personal information indefinitely, ‘just in case’ it might be useful in the future.

If you are using a third party, you **should** be clear in your contractual agreement about what personal information they retain and for how long. Third-party providers **must** be limited in the information they retain to what is necessary. This will vary depending on the service they provide. For example, a third-party moderation service that provides an AI-based classification tool may not need to retain information for longer than it takes to analyse the content. However, a third-party provider that is managing user appeals may need to keep information for longer to allow them to deliver their service.

You **should** review your retention periods regularly, and erase or anonymise personal information when you no longer need it.

You may also have to follow other laws that set out how long you need to keep certain information for.

Further reading

- Principle (e): Storage limitation

How do we ensure the security of personal information?

Data protection law requires you to process personal information securely, using appropriate technical and organisational measures but it does not define what measures to use. This is the ‘security principle’.

You **must** put in place technical and organisational measures to ensure your level of security is appropriate to the risk of using personal information. You **must** consider:

- the state of the art;
- costs of implementation; and
- the nature, scope, context and purpose of your processing.

If you plan to use a third-party moderation provider, acting as a data processor, you **must** choose one that provides sufficient guarantees about its security measures.

Further reading

- Principle (f): Integrity and confidentiality (security)
- A guide to data security

Who is the controller in our content moderation systems?

Your content moderation systems may involve different organisations. For example, you may use one or more third-party providers to support your moderation processes.

You **must** be clear about the roles that you and each party have. This depends on which of you is a data controller or processor, or whether you are joint controllers.

You are the controller, if you've made decisions about the purposes and means of any content moderation processing. You have overall responsibility for complying with data protection law. This is because you've decided the "why" and "how" of processing your users' personal information, for example:

- what you intend content moderation to achieve;
- what personal information you need for this purpose;
- what content moderation tools are involved;
- how long you'll keep the personal information for; and
- whether you engage another party to undertake the processing for you.

If you engage another party to undertake processing for you and they only act on your behalf and on your instructions, they will be your processor.

Processors can make certain technical decisions about how to process personal information, for example:

- which IT systems and methods to use in the content moderation;
- how to store the personal information in these systems;
- which security measures to apply to the personal information; and
- how to retrieve, transfer, delete and dispose of the personal information.

Example

An online service decides to use a third party that provides a content moderation tool. The tool analyses user-generated content to classify whether it violates the service's content policies. This involves processing the users' personal information.

The online service is responsible for deciding both why and how the information is processed. The third party acts only on the service's instructions, which include the specific content policies the tool has to classify against.

However, the third party uses its own expertise to carry out the moderation. It takes decisions about storing the personal information and how it transfers the results of its moderation actions back to the service.

In this case, the online service is the controller and the third-party moderation provider is the processor.

You **must** ensure your processor only carries out processing according to your instructions. If a processor acts outside of these instructions, it is processing your users' personal information for its own purposes. Not only does this mean it becomes a controller for that processing, but it is also acting outside of its agreement with you.

There can be some types of content moderation that are more complex, such as those that involve AI or several different entities. This can make it more complicated to allocate appropriate roles and responsibilities.

Further reading

- Controllers and processors
- Contracts and liabilities between controllers and processors

What data protection rights do people have?

People have a range of rights under data protection legislation. You **must** ensure that they are able to exercise their data protection rights. This also helps you comply with other data protection requirements, such as the principles.

The lawful basis you use for your processing can affect which rights are available to people.

This section considers people's data protection rights that may be most relevant to your content moderation processes.

- Can people request access to their personal information used in content moderation?
- Can people request that we rectify their personal information used in content moderation?

Can people request access to their personal information used in content moderation?

Yes. People have the right to access the personal information you use and generate through your moderation activities. You **must** provide this information if a user makes a subject access request (SAR), unless an exemption applies.

In content moderation, depending on the nature of the SAR, you may need to provide users with:

- confirmation that you are using their personal information to carry out content moderation;
- copies of the personal information you are using; and
- copies of personal information generated in your moderation systems (eg information on the content moderation outcome or action you've taken about the user or their content).

People are only entitled to their own personal information. This means that you do not need to provide additional information that is not personal information (eg confidential commercial information).

It may be challenging to respond to SARs if your systems use large amounts of information or the information contains details of other users. You **should** store the information in a way that makes it quick and easy for you to locate it.

If you are using third-party moderation providers, you **should** ensure that any personal information they use or generate is readily retrievable. You **should** factor this in when choosing your third-party moderator(s).

Can people request that we rectify their personal information used in content moderation?

Yes. People have the right to have inaccurate personal information rectified. If you receive a request for rectification, you **must** take reasonable steps to satisfy yourself that the information is accurate and to rectify it, if necessary. You **should** take into account the arguments and evidence provided by the person whose information you are using.

This is closely linked to the accuracy principle. However, even if you took steps to ensure personal information was accurate when you gathered or generated it, the right to rectification means you **must** reconsider its accuracy, if a person makes a request.

For example, your content moderation system may generate personal information about a user that states that they posted a piece of content that breaches your content policies. Even if you took steps to ensure your systems were functioning as intended, you **must** ensure the outcome you recorded about that user is accurate. You **should** consider any argument the user puts forward about that decision, and you **must** rectify the information, if needed.

Further reading

- Guidance on individual rights
- Right of access
- Right to rectification
- A guide to the data protection exemptions

How do we share information about content moderation?

There may be situations where you want to share personal information with other organisations. This may include information used in, or generated by, your content moderation processes.

Before sharing any personal information with other organisations, you **must** consider whether:

- it is necessary; or
- you can achieve your intended purpose without sharing personal information.

For example, you may want to share information with a research organisation to monitor trends in how people share certain content. In this case, it is unlikely that you need to share personal information about the users who are distributing the content.

If you need to share personal information with another organisation, you **must** identify a lawful basis. If you are disclosing special category information or criminal offence information, you **must** identify additional conditions for processing. (See section on 'How do we carry out content moderation lawfully?' for more information.)

You **should** put in place a data sharing agreement to:

- set out why you are sharing the information;
- cover what happens to the information at each stage; and
- set standards for sharing.

This helps all the parties involved in sharing the information to be clear about their roles and responsibilities.

Our data sharing code of practice provides guidance about how to share personal information with other organisations. It includes information about sharing information in an emergency and sharing information with law enforcement agencies.

The government intends to publish secondary legislation under the OSA about reports made to the NCA about CSEA content. This may include provisions about the information that you need to include in these reports. We plan to publish further data protection guidance on the requirement to report CSEA content to the NCA under section 66 of the OSA.

Further reading

- Data sharing: a code of practice
- Data sharing agreements

What do we need to consider if we transfer people’s personal information outside the UK?

Data protection law contains rules about transferring personal information to receivers located outside the UK. We refer to these as restricted transfers.

You **must** ensure that one of the following apply:

- the transfer is covered by ‘adequacy regulations’. This means that the country you plan to send the information to has ‘adequate’ protection for people’s personal information;
- there are appropriate safeguards in place. This means that if there are no UK adequacy regulations in place for that country, you **must** do an assessment to check that relevant UK GDPR protections are not undermined for people whose information is transferred; or
- the transfer is covered by an exception.

Further reading

- A guide to international transfers

What if we use automated decision-making in our content moderation?

In detail

Why is this important?

You **must** determine whether your content moderation involves solely automated decisions that have legal or similarly significant effects on people. You **should** assess this for each stage of your content moderation workflow. This is due to restrictions in Article 22 of the UK GDPR on when you can carry this out.

When does content moderation involve solely automated decision-making?

Content moderation systems can extensively use automation to support content analysis and moderation actions.

In most cases, this also means processing people's personal information because the content you are analysing is linked to a particular user's account (see the section on 'What personal data does content moderation involve?').

This can involve systems making solely automated decisions. "Solely automated" decisions are those taken without any meaningful human involvement. For example, automating the decision about whether a piece of content breaches your content policies and what type of moderation action follows afterwards.

This is particularly likely to be the case if the system you use is an AI-based content moderation tool used to classify and take action on content without a human being involved in those decisions.

Example

A service deploys an AI-based content classification system that automatically removes all content that scores above a certain confidence score in a particular category of prohibited content. For example, the system removes all content classified as 'violence' that has a confidence score of X% or greater.

The system decides whether the content meets this classification, and where it does, removes the content.

As there is no human involvement, this is solely automated.

However, not all content moderation involves solely automated decision-making. For example, this may apply to systems that use exact database matching tools. These can compare user-generated content to a database of known prohibited material that has been determined as prohibited by humans. Content that is found to be a match against this database is typically removed from the service.

These types of decisions won't necessarily be solely automated, because the moderation tool is operating according to specific, pre-defined parameters representing things that humans have already decided on. The tool isn't making a decision based on an analysis of the likelihood of something happening, unlike with classification tools.

Generally, if you intend a moderation system to go beyond exact matches of pre-defined content, then it is more likely to be making solely automated decisions. For example, it analyses additional information and makes its own predictions based on context and circumstances, such as perceptual hash matching or machine learning classification tools.

Example

A content moderation tool detects and removes links to known child sexual abuse material (CSAM).

This involves exact matching against a pre-defined list of URLs where CSAM is present. Humans have determined that these links contain such material and have added them to the database.

The decision about the nature of the content and the action to remove it is taken before the system operates and it only functions according to these parameters. In this sense, the system is not making decisions, even though it's operating automatically.

Since this is not solely automated decision-making, it does not fall under Article 22. However, the service using this tool needs to make sure it complies with all the other requirements of data protection law set out in this guidance.

When do solely automated content moderation decisions have a legal or similarly significant effect?

A 'legal effect' is something that affects someone's legal status or their legal rights. A 'similarly significant effect' is something that has an equivalent impact on someone's circumstances, behaviour or choices.

Examples of legal and similarly significant effects include decisions that:

- affect someone's financial circumstances; or
- lead to someone being excluded or discriminated against;

The impact of solely automated content moderation decisions can depend on the person, the service, and how that person uses that service. Understanding the full context in which automated decisions take place will help you identify whether Article 22 applies.

You **must** determine whether solely automated decisions taken in your content moderation systems are going to have a legal or similarly significant effect. See our guidance on automated decision-making and profiling for more information about what types of decision have a legal or similarly significant effect. If this is the case, then you need to consider the Article 22 exceptions (see next section).

Example

A video sharing service uses a solely automated moderation system that results in a user’s video being removed from the service.

The moderation action is taken as a result of a solely automated analysis that classifies the video as violating the service’s content policies.

The service is making a solely automated decision about that particular user based on analysis of their personal information.

The user is a content creator and revenue from video content is their primary source of income. Removal of the video has a significant impact on their income.

This is a solely automated decision based on the user’s personal information that has a legal or similarly significant effect on the user.

The service has a mechanism to identify which of its solely automated content moderation decisions have a legal or similarly significant effect on its users. Therefore it can determine which decisions Article 22 applies to.

The service identifies a relevant Article 22 exception for these decisions. It also implements the required safeguards and provides users with the required information about the decision-making (see next section for more information).

Further reading

- Automated decision-making and profiling including the specific section on ‘What types of decision have a legal or similarly significant effect?’

What do we need to do when Article 22 applies to our content moderation decision-making?

Consider what exception applies

Article 22 means that you **must** only take solely automated decisions that have legal or similarly significant effects if they are:

- authorised by domestic law;
- necessary for a contract; or
- based on a person’s explicit consent.

Where the decision is authorised by law

This exception applies where domestic law (including under the OSA and accompanying codes of practice) authorises solely automated decision-making with legal or similarly significant effects. But only where the law contains suitable measures to safeguard a user’s rights, freedoms and legitimate interests.

It is your responsibility to determine whether this exemption applies.

You **should** document and be able to justify which part of the legislation authorises your use of solely automated decision-making.

If you're carrying out solely automated decision-making under this exception, you **must** also comply with the requirements of Section 14 of the DPA 2018. This means you **must**:

- tell people that you've made the decision as soon as reasonably practicable; and
- be prepared for any request they may make for you to reconsider the decision, or take a new one that's not solely automated.

If someone does request that you reconsider, you **must** also:

- consider the request and any other relevant information the person provides;
- comply with the request; and
- inform the person, in writing, of the steps you've taken and the outcome.

Where the decision is necessary for a contract

This exception may apply if you’re carrying out solely automated decision-making that’s necessary to perform the contract between you and your users. For example, if you are using solely automated decision-making to enforce the terms of service that your users sign up to.

You **must** ensure that your processing is necessary for the performance of the contract. This doesn’t mean it must be absolutely essential, but it must be more than just useful.

Where the decision is based on someone’s explicit consent

This exception applies if you have explicit consent from someone to carry out solely automated decision-making based on their personal information.

It is unlikely that explicit consent exception applies to content moderation because it is unlikely to be freely given. In addition, it may be impractical for you to gather explicit consent from users.

Provide users with transparency about decisions

If your content moderation involves solely automated decision-making with legal and similarly significant effects, then you **must** proactively tell your users about this. You **must**:

- say that you're making these types of decisions;
- give them meaningful information about the logic involved in any decisions your system makes; and
- tell them about the significance and envisaged consequences the decisions may have.

For example, you **could** include this information in your privacy policy or terms of service.

You **must** also provide this information to any user that makes a SAR to you.

The OSA requires regulated services to set out in their terms of service if they are using 'proactive technology' to comply with their online safety duties. Services are also required to explain the kind of proactive technology they use, when they use it, and how it works. Complying with this duty may help you provide the transparency to users that UK GDPR requires. However, you **must** provide the necessary transparency for data protection law.

Implement appropriate safeguards

If you're relying on the contract or explicit consent exceptions, you **must** implement appropriate safeguards to protect people's rights, freedoms and legitimate interests, including enabling people to:

- obtain human intervention;
- express their point of view; and
- contest the decision.

You **should** have an appeals process for content moderation decisions that users can easily use, understand and find on your service.

Under the OSA, all user-to-user services have a duty to operate complaints processes, including for users who have generated, uploaded or shared content on the service. Among other things, they should be allowed to complain if their content is taken down (or if they are given a warning, suspended or banned from using the service) on the basis that their content is illegal. There are also complaints obligations for services likely to be accessed by children and category 1 services.

Where you are complying with your OSA complaints duties, these processes may help you provide the safeguards that Article 22 of the UK GDPR requires. In particular, ensuring your users can contest a content moderation decision. However, you **must** implement the appropriate safeguards for data protection law.

What about special category information?

You **must** also consider whether your solely automated decision-making is likely to involve special category information.

You **must not** base your decisions on special category information unless you:

- have explicit consent; or

- can meet the substantial public interest condition in Article 9.

In addition, you **must** implement safeguards to protect users' rights and freedoms, and legitimate interests.

As noted above, you are unlikely to seek explicit consent for your content moderation processing. This means that if you intend to process special category information, you **must** consider the substantial public interest condition. (See the section on 'What if our content moderation involves special category information?' for more information.)

Further reading

- Automated decision-making and profiling
- Rights related to automated decision-making including profiling
- Consent (including explicit consent)
- Guidance on AI and data protection including the section on 'What is the impact of Article 22 of the UK GDPR on fairness?'

What if Article 22 does not apply?

If Article 22 doesn't apply to your content moderation processing, you **must** still comply with data protection law and ensure that users can exercise their rights.

You **could** tell people about any automated decision-making your content moderation involves, even if it has meaningful human involvement.

You **should** tell people what information you're using and where it came from. This helps you be more transparent, particularly if your processing won't necessarily be obvious to people.

Further reading

- What if Article 22 doesn't apply to our processing?

Glossary

Artificial intelligence

An umbrella term for a range of algorithm-based technologies that solve complex tasks by carrying out functions that previously required human thinking.

Anonymisation

The techniques and approaches applied to personal information to render it anonymous.

Content classification

Automated analysis of user-generated content to assess whether it is likely to breach a service's content policies. This often involves the use of AI-based technologies. Classification systems usually assign a degree of confidence to their assessment of a piece of content.

Content moderation

The analysis of user-generated content to assess whether it meets certain standards and any action a service takes as a result of this analysis. For example, removing the content or banning a user from accessing the service.

Content removal

Action taken to remove content from a service or prevent it from being published.

Database matching

Automated analysis of user-generated content to check whether it matches an internal or external database of known prohibited content.

Feature blocking

Action taken by a service to restrict a user's access to certain features of the service, either temporarily or permanently.

Hash

A fixed length value summarising a file or message contents.

Hash matching

A technique where a hash of a file is compared with a database of other hash functions. Online services can use hash matching to detect known illegal content.

There are different types of hash matching used in content moderation. Cryptographic hash matching is used to identify exact matches of content. Perceptual hashing is used to determine whether pieces of content are similar to each other.

Moderation action

Any action that a service takes on a piece of content or a user's account after the content has been analysed.

Ofcom

Ofcom is the regulator for the OSA. It is responsible for implementing the regime and supervising and enforcing the online safety duties.

OSA

The Online Safety Act 2023. See Online Safety Act 2023 (legislation.gov.uk).

Pseudonymisation

Defined in the UK GDPR as "..processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".

Service bans

Action taken by a service to ban users from accessing the service, either temporarily or permanently.

Third-party moderation services

Organisations that provide content moderation services. This can include both content moderation technologies and human moderation.

User-generated content

Defined in the OSA as "content that is (i) generated directly on a service by a user of the service, or (ii) uploaded to or shared on a service by a user of the service, and; that may be encountered by another user, or other users, of the service by means of the service".

User-to-user service

Defined in the OSA as "an internet service by means of which content that is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, may be encountered by another user, or other users, of the service".

Visibility reduction

A range of actions that services may take to reduce the visibility of content. For example, preventing content from being recommended or making content appear less prominently in users’ news feeds.