Content moderation and data protection guidance impact assessment

February 2024

ico.
Information Commissioner's Office

Contents

1. Background and context

The ICO has an 'Impact Assessment Framework' (IA Framework) which sets out when we do and don't carry out impact assessments (IAs). Here we explore the application of the Framework to our guidance on content moderation.

1.1. Background

The ICO committed to publishing guidance on online safety technologies in its 2022 Joint Statement with Ofcom on Online Safety and Data Protection, as part of our ongoing work to ensure regulatory coherence between the two regimes. We committed to preparing guidance that would be separate from but aligned with the codes of practice and guidance that Ofcom are required to produce under the new Online Safety Act (OSA).

1.1.1. Online Safety Act

The OSA received royal assent in October 2023. It requires providers of regulated user-to-user and search services to have certain duties of care to protect users from illegal content. If a service is likely to be accessed by children, it also sets out duties for the protection of children.

The Department for Culture, Media and Sport (DCMS)[1] produced an impact assessment for the OSA. The impacts considered in this document are distinct from those covered in the IA for the Act.

Ofcom, the regulator for online safety, has set out their plan to implement these new rules from the OSA. In three phases, Ofcom plans to give guidance and set out codes of practice on how in-scope companies can comply with their duties.

1.1.2. Call for views

The ICO conducted a call for views on data protection and content moderation from 26 April to 9 June 2023. The call for views received a total of 15 responses, of which, four covered impacts from organisations involved in content moderation. A summary of responses will be published in Spring 2024. Informed by the views collected, a programme of direct stakeholder engagement, and working in tandem with Ofcom to ensure alignment, the ICO has developed guidance on content moderation and data protection. This is the first in a series of planned products on online safety technologies.

[1] Note: responsibility for delivering the Online Safety Bill moved from DCMS to a new department, the Department for Science, Innovation & Technology (DSIT), after machinery-of-government changes in February 2023. This was after the impact assessment was published in January 2023.

1.2. The guidance and impact

The guidance sets out how organisations deploying content moderation processes or providing content moderation services can comply with UK General Data Protection Regulation (GDPR) and the Data Protection Act (DPA) 2018. It is relevant for all organisations who are carrying out or considering carrying out content moderation and providers of content moderation products and services, but our primary audience is those doing this to meet their obligations under the Online Safety Act 2023 (OSA).

The guidance is not intended to be a comprehensive guide to compliance but it does aim to deliver significant impacts on the interests of data subjects, data controllers, and data processors by clarifying how data protection applies to content moderation. As stated in our IA Framework criteria, we are more likely to carry out an IA where there are likely to be significant impacts on these groups.

In terms of proportionality, the guidance sets out the requirements of existing data protection law where personal information is processed in content moderation, thus the guidance does not place additional obligations on organisations above those in existing legislation.

In these circumstances, as per the principles set out in our IA Framework an impact assessment summary table approach is proportionate here, as set out below. The assessment of impacts is presented in Section 2 and the theory of change is provided in Figure 1 below.

2. Application of our impact assessment approach

As outlined in our IA Framework, IAs include the following six elements:

1. problem definition;
2. rationale for intervention;
3. identification of alternatives;
4. description of the regulatory proposal;
5. analysis of benefits and costs; and
6. setting out the proposed monitoring and evaluation needs.

Steps 1 – 4 are covered in Section 2.1, with step 5 addressed in Section 2.2 and step 6 in Section 2.3.

2.1. From problem definition to detail of the intervention

The table below provides more detail on the journey from problem identification to the proposed intervention. It covers the market failures and data protection harms we have identified, the groups affected and the options we have considered.

Table 1: Impact assessment, steps 1-4

1: Problem definition	The OSA received royal assent in October 2023. The OSA gives Ofcom the authority to regulate online content on regulated user-to-user services and search services. It focuses on illegal content, as well as content that is harmful to children. The regulation of online content can interact with data privacy concerns in important ways. For example, decisions around the use of biometric technology to establish the age of users is likely to require consideration of how this data is collected and stored, and who has access to it. Organisations in scope of the OSA may require greater regulatory certainty about how data protection law will apply to online safety technologies, including content moderation. This was supported by the call for views which found general uncertainty among organisations about how the data protection regime applies to content moderation processes, particularly in the context of the OSA.

2: Rationale for intervention

Market failures

The guidance mitigates against potential market failures resulting from a lack of clarity on how to conduct content moderation while respecting data protection principles. This can present as inefficiently high costs: without the guidance, organisations could be incurring higher compliance costs. This could include seeking legal advice or the costs associated with legal or regulatory action. Another potential market failure is negative externality: without ICO guidance there is a risk that online safety objectives will take priority over privacy considerations. There may also be information failures, where organisations are not clear about the privacy information that they should provide to users of their services. This would erode users' privacy and information rights.

Policy and legal context

In the 2022 Joint Statement with Ofcom on Online Safety and Data Protection, the ICO commits to "prepare guidance on data protection expectations for online services deploying safety technologies (e.g. age assurance, content moderation) and will consult Ofcom, amongst others, in its preparation".

The ICO intervention aligns Ofcom's regulatory duties from the Online Safety Act 2023 with the ICO's regulatory approach to data protection laws, making it easier and less costly for organisations to comply.

Data protection harms

There is potential for data protection harms resulting from content moderation that doesn't comply with data protection principles. This could include, in the immediate term, loss of control of personal data. For example, an organisation carrying out content moderation could unexpectedly and unfairly process or share personal data as part of the process. There could then be medium or longer term harms that occur as a result. These include: financial and psychological harms, chilling effects, and discrimination on the basis of a moderation system's outputs. For example, content moderation that uses automated processes that are susceptible to bias and discrimination could lead to loss of income if a person relied upon their content to generate income.

Affected groups

The main actors and groups expected to be impacted by this guidance include:

- Organisations using or providing content moderation products and services. These can be split into three groups:
    - organisations who use content moderation solutions that they develop in-house;
    - organisations that use content moderation solutions and services supplied by a third party; and
    - third-party service providers who develop and supply content moderation solutions and services for others to use.
- Other regulators and public bodies, and in particular Ofcom, who may refer to our guidance as part of their implementation of the online safety regime.
- Individual users of regulated services whose data is processed by content moderation technology.
- The wider population who may experience knock on effects such as the societal impacts of reductions in data protection harms.
- The supply chain of the organisations who carry out content moderation.
- Wider society.

Organisations that use content moderation services span across many sectors, and include organisations that are more innovation focused. The call for views identified that content moderation can bring significant costs, and therefore organisations can be more sensitive to changes in this area.

Summary

The potential for market failures, the nature of policy alignment, the potential for data protection harms, and the scale of possible cohorts affected present a strong rationale for intervention by the ICO.

3: Options appraisal

The options considered are formed around the existing commitment in the 2022 Joint Statement with Ofcom on Online Safety and Data Protection. The options considered provide a good sense of the implications of alternative approaches and demonstrate why the ICO decided on the preferred option. The options considered include:

	1. Revisit: revisit the original commitment made in the joint statement and decide whether it is still appropriate or requires adjustment. 2. Preferred: High-level guidance setting out the ICO’s preliminary data protection and privacy expectations for online content moderation, and providing practical examples, with plans for further work as the policy area develops. 3. Do more: More extensive guidance discussing in depth how data protection law applies when developing or using content moderation. Option 2 was identified as the preferred option. This is because it provides some degree of clarity for a wide variety of stakeholders, whilst still allowing the necessary flexibility for our policy positions to develop during the early stages of Ofcom’s policy and guidance development.
4: Detail of proposed intervention	The intervention is focused on developing guidance that sets out how organisations deploying content moderation processes or providing content moderation services can comply with data protection law. The guidance aims to ensure regulatory coherence between data protection legislation and Ofcom’s new duties from the OSA. The changes that this guidance aims to bring about are: • capitalising on the opportunity to improve and maintain compliance with data protection legislation alongside the implementation of the new online safety regime; • improved confidence among developers and users in applying data protection legislation to their content moderation solutions; • reduced compliance costs for organisations; • more efficient, effective and competitive organisations; • positive impact on individuals’ rights and freedoms, including data privacy and freedom of expression; and • a reduction in data protection harms. The ICO plans to keep this guidance under review and update it in due course to reflect Ofcom’s final online safety codes of practice and guidance.

Source: ICO Economic Analysis.

2.2. Cost-benefit analysis

The costs and benefits of the intervention have been identified, as far as is possible and proportionate. Below is an overview of the primary costs and benefits we have considered. This should not be viewed as exhaustive or hierarchical.

There is limited quantitative data and the analysis relies heavily on qualitative information which increases the uncertainty of the assessment. Bearing in mind these caveats, our overall assessment of the intervention suggests that the benefits are likely to outweigh the costs. This is largely due to the guidance itself being unlikely to impose significant costs over and above the existing legislation.

Table 2: Cost-benefit analysis overview

Affected group	Benefits	Costs	Scale of affected population
Organisations using or providing content moderation products and services	• improved understanding of how to comply with data protection law when developing or carrying out content moderation; • reduction in compliance costs; • improved public confidence in organisations with increased data protection compliance.	• initial familiarisation costs[2] with the new guidance. We estimate that the familiarisation cost is approximately £80 per organisation.	The OSA IA estimates 21,500 organisations could be affected by the Act. This could be used as an indicative proxy. Assuming up to 65% engagement with guidance[3] we could estimate that 13,975 organisations would be affected.
People who use regulated services and whose data is processed by content moderation technology	• access to better and more compliant services; • improved ability to exercise data protection rights such as right to be informed or right not to be subject to a decision based solely on automated processing; • improved ability to exercise non data protection rights such as the right to freedom of expression; • reduction in data protection harms.		It is difficult to estimate the number of potential users but given the widespread use of services where content moderation could be applied, the total UK population of 67 million people (estimated by the ONS) could be applied as an upper estimate.

[2] Familiarisation costs are the costs associated with reading and becoming familiar with new or revised guidance. We calculate these as administrative costs associated with an individual at manager, director or senior official level reading the document. See Business Impact Target guidance or Annex A of our previous IA for the Data protection and journalism code for more information on the approach.

[3] The Business Perceptions Survey 2022 estimates that, across sectors, the share of businesses that use guidance is 65%.

ICO	• reduced number of complaints resulting from non-compliant content moderation; • ability to allocate more resources to focus on improving compliance.	• upfront resource costs of production, awareness raising, and monitoring of the guidance.	This group is wholly represented by the ICO, however it is worth noting here the potential to affect Ofcom and other regulators.
Wider society (including people and organisations)	• reduction in societal costs associated with data protection harms; • economic and societal benefits of more efficient, effective and competitive organisations.		As with the "people" effected group, the total UK population could be used as an upper end estimate of the number of people that could be affected by societal impacts. Given the difficulty in estimating a total number of organisations directly affected by content moderation, it would not be possible to provide a robust estimate of those indirectly affected.

Source: ICO Economic Analysis.

2.3. Monitoring and evaluation

Finally, as per our IA Framework, we consider monitoring and evaluation. In line with best practice and organisational standards, we will put in place an appropriate and proportionate review structure. This could include:

- feedback from organisations on the guidance;
- engagement figures that monitor how many times the guidance is viewed; and
- working with Ofcom to seek some alignment and complementarity between our monitoring and evaluation activities.

3. Theory of change

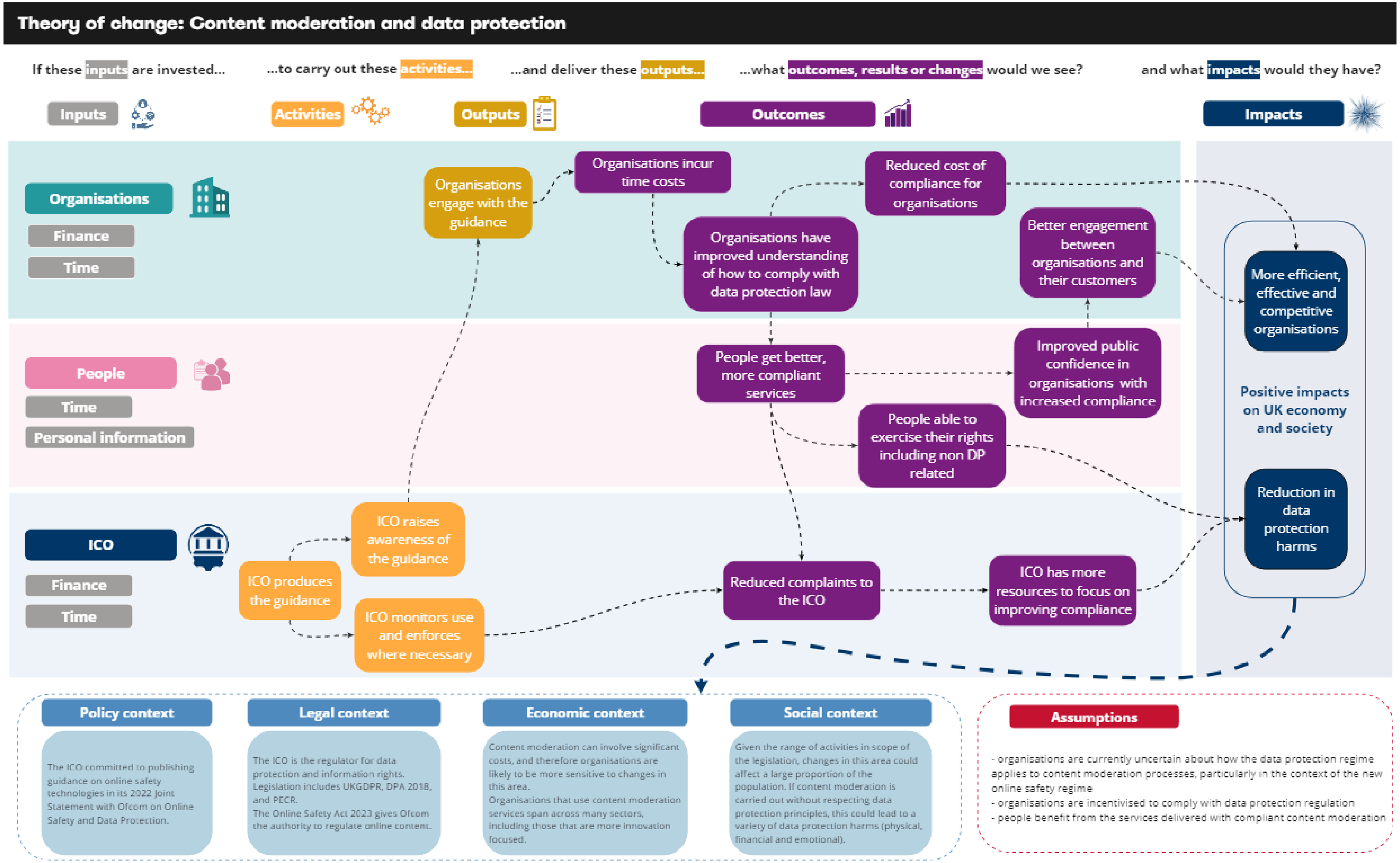
Figure 1 below illustrates our theory of change. A theory of change is a systematic approach used in intervention design and evaluation that enables us to produce a visual and/or narrative representation of how and why an intervention is expected to work to drive change. It outlines the causal pathways and linkages between inputs, activities, outputs, outcomes and impacts.

This theory of change was developed in collaboration with the project delivery team and is intended as a visual representation of the causal pathways for the benefits and costs considered in Table 1.

Our theory of change shows the link between the affected groups identified (organisations, people, the ICO) and the intended impact of our proposed intervention for organisations using or providing content moderation products and services.

For example our theory of change shows that the ICO will produce the guidance and raise awareness of it, while also monitoring its use (activities). Organisations will engage with the guidance (output). This will initially involve familiarisation costs, and result in improved understanding of the regulatory environment, which will reduce their compliance costs. Greater understanding will then lead to an increased level of data protection built into content moderation (outcomes), reducing data protection harms and leading to higher levels of trust and better engagement between organisations and their customers (impacts).

Figure 1: Content moderation guidance theory of change



Theory of change: Content moderation and data protection

If these inputs are invested... ...to carry out these activities... ...and deliver these outputs... ...what outcomes, results or changes would we see? and what impacts would they have?

Inputs

Activities

Outputs

Outcomes

Impacts

Organisations

Finance

Time

People

Time

Personal information

ICO

Finance

Time

Organisations engage with the guidance

Organisations incur time costs

Reduced cost of compliance for organisations

Organisations have improved understanding of how to comply with data protection law

Better engagement between organisations and their customers

People get better, more compliant services

Improved public confidence in organisations with increased compliance

People able to exercise their rights including non DP related

ICO produces the guidance

ICO raises awareness of the guidance

ICO monitors use and enforces where necessary

Reduced complaints to the ICO

ICO has more resources to focus on improving compliance

More efficient, effective and competitive organisations

Positive impacts on UK economy and society

Reduction in data protection harms

Policy context

The ICO committed to publishing guidance on online safety technologies in its 2022 Joint Statement with Ofcom on Online Safety and Data Protection.

Legal context

The ICO is the regulator for data protection and information rights. Legislation includes UKGDPR, DPA 2018, and PECR.

The Online Safety Act 2023 gives Ofcom the authority to regulate online content.

Economic context

Content moderation can involve significant costs, and therefore organisations are likely to be more sensitive to changes in this area.

Organisations that use content moderation services span across many sectors, including those that are more innovation focused.

Social context

Given the range of activities in scope of the legislation, changes in this area could affect a large proportion of the population. If content moderation is carried out without respecting data protection principles, this could lead to a variety of data protection harms (physical, financial and emotional).

Assumptions

- organisations are currently uncertain about how the data protection regime applies to content moderation processes, particularly in the context of the new online safety regime
- organisations are incentivised to comply with data protection regulation
- people benefit from the services delivered with compliant content moderation

Source: ICO Economic Analysis.