SAFETY BY DESIGN CODE OF PRACTICE

May 2026

PROF. LORNA WOODS OBE

Introduction

Safety by design has emerged as a central principle in digital regulation, reflecting a shift toward tech accountability that requires digital services to assess and mitigate risks to users from the earliest stages of product development and throughout the entire lifecycle of the product or service.

While there is broad consensus that safety by design involves proactively building protections into systems and designing out risks to ensure user safety, there remains less clarity around the specific measures platforms must adopt and how these principles translate into operational practice. This code of practice seeks to address that gap by providing a practical overview of safety by design, within the framework of the Online Safety Act and Ofcom's existing codes and guidance.

THIS MODEL CODE HAS BEEN CO-PRODUCED WITH THE FOLLOWING ORGANISATIONS:

Online Safety Act Network

5Rights Foundation

End Violence Against Women Coalition

Glitch

Internet Watch Foundation

Refuge

NSPCC

Molly Rose Foundation

FlippGen

Professor Lorna Woods OBE

THIS CODE IS SUPPORTED BY:

Age Check Certification Scheme

Anti-Bullying Alliance

Antisemitism Policy Trust

Centre to End All Sexual Exploitation(CEASE)

Center for Countering Digital Hate (CCDH)

Centre for Protecting Women Online, The Open University

Centre of expertise on child sexual abuse

Children & Young People's Mental Health Coalition

Childnet

Clean Up The Internet

Everyone's Invited

Gender + Tech Research Lab, University College London

HOPE not hate

Institute for Strategic Dialogue (ISD)

Marie Collins Foundation

Minderoo Centre for Technology and Democracy

National Children's Bureau

Plan International UK

Samaritans

Save the Children UK

Shout Out UK

Survivors Against Terror

Tender

Suzy Lamplugh Trust

SWGfL

Thomas William Parfett Foundation

The Jo Cox Foundation

YoungMinds

White Ribbon UK

Adele Zeynep Walton,
Online safety campaigner and author of Logging Off

Professor Clare McGlynn,
Durham University

Professor Lisa Sugiura,
University of Portsmouth

Professor Sonia Livingstone,
Digital Futures for Children

Steve Wood, Senior Visiting Fellow,
Digital Futures for Children (Former UK Deputy Information Commissioner)

Table of Contents

CONTEXT 5

SAFETY BY DESIGN CODE OF PRACTICE 6

What is this code? 6

Who is it for? 6

What is safety by design? 6

Where does this sit within the legislative framework? 8

Our approach 9

THE CODE 11

Part 1 **Governance** 11

Part 2 **Preparation and development** 13

Purpose 13

Processes and systems services should have in place 13

Prohibited practices 15

Part 3 **Deployment and monitoring** 20

Terms of Service 20

Robust age assurance mechanisms 22

Moderation 24

User tools 27

Transparency 29

Redress mechanisms 31

Part 4 **Retirement and decommissioning** 33

Retirement and decommissioning 33

ENDNOTES 35

Context

Increasingly we are seeing safety by design referenced in relation to the digital space. Recent high-profile cases against Meta and YouTube in the US relating to the addictive design of their platforms have highlighted the role of design choices on user safety and wellbeing[1]. Yet despite clear evidence that design choices made by social media and tech platforms can be harmful[2][3][4], civil society organisations have highlighted a lack of clarity around what safety by design means in practice and called for safety by design to be more clearly mandated to ensure platforms do not continue to put users in harm's way[5][6][7]. Indeed, whilst Ofcom have not interpreted the OSA's reference to safety by design[8] clearly or comprehensively in their codes of practice, at a recent NSPCC event Melanie Dawes, the Chief Executive of Ofcom, said unequivocally that for tech platforms "Safety by design is not optional. It is the law."[9]

Beyond the UK, international principles and legislation, such as the EU Digital Services Act[10] and guidance from the Australian e-safety Commissioner emphasise accountability, transparency, the assessment and mitigation of risk, and the protection of fundamental rights for users[11]. The World Economic Forum's Global Coalition for Digital Safety advocates for digital safety design interventions that draw from "safety by design principles and best practices to provide a resource to assist companies in effectively identifying and reducing digital risks, preventing harm and promoting trust and safety"[12]. Furthermore, a recent statement from Mama Fatima Singhateh, the UN Special Rapporteur on the sale, sexual exploitation and sexual abuse of children, highlighted that "Despite existing national, regional and international regulations and guidelines calling for the design and development of digital products with the highest level of privacy, safety and security for children, enforcement remains inconsistent, with limited oversight and liability frameworks. Many companies fail to implement safety by design approaches, robust age verification and algorithmic transparency.[13]" A safety-by-design approach as outlined in our code of practice enables platforms to align with these global expectations.

Safety By Design Code of Practice

WHAT IS THIS CODE?

Safety by design has become a core tenet in digital regulation. While there is agreement on what it means in broad terms, there is less certainty about what is specifically required and how that might work in practice. This code aims to provide an overview of safety by design set within the framework of the Online Safety Act and Ofcom's existing codes and guidance. Our code takes a more ambitious approach than Ofcom has to date, however its positioning within the framework of the OSA means that, with just a few technical amendments, Ofcom could adopt this swiftly.

WHO IS IT FOR?

This Code of Practice provides detailed guidance for all tech companies to help them understand safety by design and how safety by design principles might be applied in the context of digital services (including but not limited to services currently within scope of the Online Safety Act). It also serves as a template for adoption by the Government and Ofcom as a model for delivering on the Act's requirement that regulated services are "safe by design" and realising, with no further delay, Parliament's ambitious intent when it passed it. This document identifies key principles as well as examples of good practice.

WHAT IS SAFETY BY DESIGN?

- **Safety must be embedded into product design decisions from the outset and retroactively. It requires:**
  - A hierarchy of control approach
  - Application across the product in its entirety
  - Consideration of safety across the product lifecycle, as opposed to merely retroactive or 'add-on' safety measures.

Hierarchy of control[14 15] means that services should seek to design out hazards from the outset. Where this is not possible, services should seek to manage and mitigate risk of harm. Remediation is a mechanism of last resort. In making design choices, services should seek solutions which best optimise outcomes for users, rather than relying on harsh trade offs.

As part of this approach, risk assessments are key and appropriate product testing essential. For instance, tech platforms should be approaching carrying out risk assessments in a holistic way, scrutinising their products through audits such as (but not limited to) algorithm quality testing, including testing for different types of bias and discrimination[16] (including intersecting biases), human rights impact assessments and

environmental safety testing that accounts for the broader impacts across the entire value chain.

Product entirety means that safety issues should be considered across the product in its entirety - this would include not just user-facing interfaces but also the backend elements and business models.

Service providers should also be assessing their business model and revenue generation strategies for a service through a safety and human rights lens. This should include consideration of: the impact that the proposed model/strategy would be likely to have on user safety, taking into account the current state of knowledge from both internal and external research on tech-facilitated harms; the impact that the proposed model/strategy would be likely to have on human rights, including privacy and freedom of expression, taking into account the current state of knowledge from both internal and external research on tech-facilitated rights abuses; and should identify changes that can be made to alleviate these risks, which could look like moving away from features being designed to maximise engagement to drive revenue, minimising the amount of data collected on users, or increasing interoperability with other services.

Safety by design should be considered when thinking about account sign up processes, content creation tools, content discovery and curation functions and tools for users to react to the content of others, as well as content moderation tools (including design of avenues for users to report content and complain). Hazards should be tackled as close to source as possible and reduce the likelihood of harm occurring or its extent or severity. Moreover, features should be considered individually but also in combination, both as regards to whether there is a hazard or risk of harm, or whether there is an appropriate mitigation/management tool. This more systemic approach to designing and testing the service links with the idea of optimisation. In making these choices service providers must take account of the severity of various risks of harm. While user controls can be valuable in supporting user choice and autonomy, users cannot be made entirely responsible for their own safety.

Product lifecycle requires consideration of safety at the product development stage as well as through ongoing monitoring and updating, and finally decommissioning. This is sometimes termed a holistic approach. This code is structured around the product lifecycle.

As numerous studies have shown[17,18], different users have different exposure to harm, and different features may produce additional or more acute risks for some users while, for the service provider, there may be a risk of privilege hazard. For a product or service to be "safe by design", it should be designed in accordance with design justice principles[19]. This includes (following Sasha Costanza-Chock[20] and the Design Justice Network[21]):

- **Consideration of how different groups have participated in the design of the product or service**
- **Consideration of how different groups may be harmed by the design of the**

product or service, and taking steps to remediate where certain groups are disproportionately harmed

- Consideration of how different groups may benefit from the design of the product or service, and how those benefits can be fairly allocated

WHERE DOES THIS SIT WITHIN THE LEGISLATIVE FRAMEWORK?

Ofcom is required to produce codes to help services fulfil their safety duties (section 41). While the Act specifically requires a code dealing with terrorism and one dealing with child sexual abuse material, s 41(3) specifies that Ofcom should prepare one or more codes proposing measures to satisfy the safety duties - and that relates to the illegal content safety duties, the safety duties relating to protection of children as well as the duties on categorised services. Ofcom could use this power to base its actions, and the safety by design code could underpin the requirements of the other codes.

The OSA explicitly references the need for user-to-user services to be 'safe by design' on the face of the Act in section 1(3). It reads;

> *Duties imposed on providers by this Act seek to secure (among other things) that services regulated by this Act are—*
>
> *(a) safe by design, and*
>
> *(b) designed and operated in such a way that—*
>
> *(i) a higher standard of protection is provided for children than for adults,*
>
> *(ii) users' rights to freedom of expression and privacy are protected, and*
>
> *(iii) transparency and accountability are provided in relation to those services*[22].

Furthermore, the Secretary of State for the Department for Science, Innovation and Technology (DSIT) sets out safety by design as a key priority in their Statement of Strategic Priorities (SSP)[23] for Online Safety making clear that Ofcom, in having regard for the SSP, should ensure platforms;

> *Embed safety by design to deliver safe online experiences for all users but especially children, tackle violence against women and girls, and work towards ensuring that there are no safe havens for illegal content and activity, including fraud, child sexual exploitation and abuse, and illegal disinformation.*

Yet despite this approach being a core tenet of the Act and both Parliament and Government's expectations from it, Ofcom do not define what a safety by design approach must look like in any of their codes of practice. Whilst their more recent work does include express reference to safety by design, such as in the VAWG guidance[24], its interpretation of "safe by design" is limited mainly to a few ex-post measures. Even their proposals on recommender algorithms seem to focus on individual items of content rather than how the algorithms are designed in the first place[25]. Furthermore, they have not taken a holistic approach to what this means in terms of the design and operation of services,

their systems and processes or their business model, even where specific risks relating to features and functionalities have been evidenced in the risk register. [26]

OUR APPROACH

The following code of practice seeks to address safety by design through the entire lifecycle, considered across the product in its entirety, and taking a hierarchy of control approach. The early design process is key to identifying and seeking to mitigate risk (e.g. illegal content and activity, harm to children and risks arising from the service's business model [including incentives to expand the user base, increase engagement, and encourage the sharing of user-generated content]). Whilst we have referenced the need for robust risk assessments to be carried out as part of a hierarchy of control approach, we have chosen to follow the structure of the OSA, which sets out separate code duties for risk assessments, supported by guidance which Ofcom has also produced[27]. As such, we refer providers to those duties in this code, whilst recognising the need for Ofcom to iteratively expand on their understanding of risk in line with emerging technologies, such as AI chatbots, and how emerging risks may show up for and affect users differently.

Risks can evolve once a service is live and interacting with real users at scale. Having anticipated potential harms and risks during the design stage, the second stage of a safety-by-design approach requires services to ensure that protective measures are effectively implemented when a product, feature or business strategy or model (e.g. user expansion) is deployed.

With regard to deployment and monitoring, Ofcom's Codes of Practice focus on mechanisms to detect, assess and respond to illegal content and content harmful to children, alongside transparency obligations about processes, enforcement and moderation. While these rules provide important minimum standards for compliance, they are largely process-oriented and, due to safe harbour provisions, specify what services must do to meet their duties, rather than how to proactively embed safety in ongoing operations.

A safety-by-design approach builds on this by treating deployment as an extension of the design process. It ensures that safety measures are fully integrated, tested in real-world conditions, monitored and adapted in response to emerging risks or unforeseen patterns of harm. This approach goes beyond Ofcom's code measures by embedding accountability, iterative evaluation and real-world effectiveness at the core of the service. This approach also supports services to apply safety-by-design retrospectively, helping them to identify unforeseen risks, correct unsafe patterns of interaction and adjust safeguards to improve outcomes. It should also ensure that product designs which, upon deployment, mean that there are risks disproportionately experienced by members of marginalised groups, or unforeseen negative consequences for the protection of human rights, are identified and changed.

During deployment and monitoring (part 2), services should embed continuous product monitoring, risk and impact assessment into all processes. This means regularly reviewing moderation systems, age assurance, user tools, and other safety features at

appropriate intervals proportionate to the level of potential risk, to ensure they work as intended and adapt to new or emerging risks. Insights from user reports, internal data, external audits, and robust scientific research should feed into ongoing improvements, helping to maintain protections and prevent safety measures from weakening over time. Audit processes should engage with members of groups most affected by risks and organisations representing them.

Effective safety-by-design requires robust governance at every stage of a service's lifecycle. Good governance should ensure that safety measures are planned, implemented, monitored, and iterated in practice. Boards, senior managers, and relevant decision-makers must have clear oversight of risks, mitigation strategies, and emerging harms, and must be able to make informed decisions based on evidence and outcomes. There must be clear allocation of accountability for safety-by-design, with clear reporting lines and sufficient mitigation of any conflicts of interest that could otherwise weaken a culture of responsible behaviour and continuous improvement. There should be independent accountability mechanisms in place to ensure that the protection of human rights is integrated effectively throughout the service lifecycle.

Decisions should be guided by ethical and rights-respecting principles, with a particular focus on children, women, marginalised groups, and other vulnerable users. Governance must consider the real-world impact of design choices on different groups, contexts and environments, including algorithmic amplification, feature changes, or content-sharing mechanisms, and ensure that these do not inadvertently increase risk to unacceptable levels. Where remediation for risks is necessary, this requires investment in context-sensitive mitigations.

Transparency is a core component of governance. Decision-making processes, evidence of effectiveness, and actions taken in response to risks should be documented, shared internally, and made accessible to regulators, researchers, and where appropriate, the public. Where children are affected, reporting must be clear and age-appropriate, showing how protections are maintained and harms are mitigated.

This governance approach aligns with IEEE 2089 guidance on age-appropriate design[28], embedding accountability, measurement, and continuous improvement into every stage of the service lifecycle.

The Code

PART 1 GOVERNANCE

01 Regulated services should have a specific policy commitment to prevent harm and to take action to ensure their service is safe by design for all users. This commitment should be endorsed by the UK leadership of the organisation and a board member, or person reporting into the board, appointed to be accountable for delivering it. The policies should be informed by specialist expertise reflecting the experience of different groups using or affected by the service. It should clearly set out the values of the regulated service.

02 Services must implement governance structures that ensure safety considerations are embedded in strategic decision-making and product development. This must include:

- **a.** Board-level oversight of safety risks and named responsible person(s), where the level of oversight is proportionate to the level of safety risk
- **b.** Clear and public allocation of responsibility for compliance with this code including by specifying a named person with overall responsibility for safety by design, as well as senior managers in relevant roles
- **c.** Regular internal review of safety performance and risk mitigation measures, conducted by personnel with sufficient resources, expertise and independence as necessary, and which engages directly with users most affected by risks
- **d.** Conduct rules for senior managers, requiring actions that promote and uphold user safety, ensure transparency and disclosure, and necessitate due regard to the best interests of the child
- **e.** Integration of safety-by-design principles throughout the product lifecycle
- **f.** Independent oversight mechanisms

COMMENTARY

Active leadership in companies is necessary to ensure safe design: ISO/IEC 27001 recognises the importance of leadership and commitment[29]. It is also important for ensuring that incentives within the company which influence design and development choices are aligned with safety objectives. We note that Ofcom's Illegal Content Codes state providers should name an individual accountable to the most senior governance body for compliance with the duties, including the reporting and complaints duties. The code with its expectation of board level oversight and named person responsibility reflects that approach, save that the responsibility is directed towards safety by design. The same person as identified under the Ofcom codes could be designated here. While not required here or by Ofcom's code, it seems sensible for service providers to undertake a mapping exercise to identify roles and departments relevant to discharge of functions of this code. The governance structures are relevant throughout the lifecycle of the product; they do not apply just to product design and development stages.

EXAMPLE

Introduced in response to the 2008 banking crisis, The Senior Managers and Certification Regime (SMCR) is a regulatory framework designed to ensure accountability and high standards within financial services.

Under the Senior Managers Regime, individuals in senior roles must be approved by regulators, pass a 'fit and proper' test, and adhere to a 'statement of responsibilities' that outlines their key obligations. The Certification Regime applies to other employees whose roles could pose significant risk to customers, requiring firms to assess that staff are able to do their job competently and safely annually.

Alongside this, the SMCR establishes conduct rules that apply to all staff, promoting integrity, diligence, cooperation with regulators, and fair treatment of customers. Additional rules for senior managers emphasise effective oversight, regulatory compliance, responsible delegation, and proactive disclosure of information of which regulators would reasonably expect notice.[30]

PART 2 PREPARATION AND DEVELOPMENT

PURPOSE

03 Before developing a service or functionality, providers of regulated services must consider and document how the purpose and business model of the service is compatible with user safety, non-discrimination and the protection of human rights, including privacy and freedom of expression. This should include consideration of the ways in which the purpose and business model of the service promotes user safety, non-discrimination and the protection of human rights.

PROCESSES AND SYSTEMS SERVICES SHOULD HAVE IN PLACE

04 **Understanding hazards and mitigation**

a. **In support of the risk assessment required by the OSA[31], services should generate risk acceptance criteria, taking into account Ofcom's risk register and relevant risk profiles and taking into account the potential impact on different groups of users. The criteria should reflect the service provider's publicly stated values. Services should develop appropriate metrics for quantification of risk and develop a baseline standard of hazards and risk of harm to inform the risk assessment. Services must engage with civil society, experts, and individuals with lived experience on this process.**

b. **Following testing, and carrying out risk assessments in accordance with Ofcom guidance, document mitigation strategies taking into account the Hierarchy of Control: Services must prioritise risk mitigation measures according to the following hierarchy. (Services must demonstrate that higher-level controls have been considered before relying on lower-level controls, and be able to justify why those higher-level controls were not usable)[32].**

- **Elimination:** Remove features or system designs that create a foreseeable risk of harm.
- **Substitution or design modification:** Redesign features to reduce the likelihood or severity of harm.
- **Engineering and technical safeguards:** Implement robust technical systems that limit or detect harmful behaviour.
- **Administrative controls:** Actionable and upheld policies, moderation systems and governance processes to manage risks.
- **User-level tools:** Mechanisms enabling users to control their experience or respond to harm.

That the design and operation of the service meets minimum standards of human rights and non-discrimination, and if it fails to do so, the service or functionality cease to be deployed until remedial action is taken

Mitigation strategy should be kept updated and respond to any findings developed under product testing or market surveillance mechanisms.

05 **Accountability:** Evidence of compliance, documented risk assessment and governance processes, mitigation strategies, details of board oversight and senior manager responsible for this documentation. Publicly available documents demonstrating appropriate consideration (and where feasible mitigation) of potential risks to children.

06 **Trust and safety teams:** Services must maintain adequately resourced and trained trust and safety teams responsible for implementing safety obligations under the code.

a. This must include a dedicated staff or team member responsible for online safety, harm prevention and moderation, with access to sufficient staffing levels to respond to reports and emerging harms in a timely manner.
b. Ongoing training, adequate compensation, welfare provisions and psychological support for staff, including where outsourced
c. Clear internal and external escalation processes for serious harms (e.g. child exploitation, threats or live events)
d. Regular reporting to senior leadership and the board on safety risks, incident trends and mitigation actions, and publication of these reports for public transparency
e. Integration into product development, ensuring safety is considered in design decisions.
f. Sufficient authority and/or independence to carry out their roles while minimising any risk of conflict of interest.
g. Sufficient expertise in how risks manifest in different contexts and for different groups

07 **Safety Testing:** Services must test existing and new features, settings, and policies for potential harm, distinguishing between harms arising from the implementation of features, settings or policies, mass take-up and those arising from malicious use of the service or feature. Safety testing must be a core input of risk assessments, and be carried out to assess the effectiveness of mitigation strategies. This includes but is not necessarily limited to pre-launch safety reviews; adversarial testing and abuse scenario simulations. Testing should be based on the risk profile of a service - identifying areas of risk and assessing the robustness of safety measures. It should also encompass known pathways to harm resulting from the use of features in combination. Safety testing must take into account and respond to how users with different characteristics experience different forms and levels of harm, including on the basis of gender, race, ethnicity, age, disability, sexual orientation, religion or belief, and other protected characteristics. Particular regard and remediation should be paid to where marginalised groups experience disproportionate levels of harm resulting from the features and functionalities of a service. Testing should be done with independent input and oversight.

PROHIBITED PRACTICES

08 **Prohibited high risk practices:** There are some features and functionalities which should not be employed at all. Services must not deploy interfaces that manipulate, pressure, or mislead users into making decisions that reduce their safety, privacy, or wellbeing. This prohibition applies to all users, with enhanced protections required where children are likely to access the service.

09 Platforms should have due regard to industry best practice standards, independent research and expertise, Ofcom's risk profiles, and experiences of users affected by risks, in defining which specific steps they ought to take to mitigate risks on their services. In putting in place such steps, platforms should also have regard to how their mitigations could have consequences for the protection of human rights and/or disproportionately affect different groups.

Below is a non-exhaustive list of examples of steps which could be put in place before service deployment to reduce commonly occurring risks to users, particularly children.

10 **Account Creation**

Regulated services should determine an appropriate approach to account security and ensure, and be able to demonstrate, that their sign-up processes have taken an appropriate and proportionate approach to the principle of "knowing your client" (KYC), both in relation to users and advertisers (though the same system need not be applied to both).

11 **Default safety settings**

a. Accounts are private by default; when the user is a child the ability of the child to switch this off should be linked to the evolving capacity of the child
b. Restricted contact from unknown adults for children, including restrictions on adults being able to search for and find child accounts as well as contact them.
c. Protections against algorithmic surfacing of harmful content for children
d. Location sharing should be off by default; and for children must require parental review, where appropriate and safe
e. Children's accounts not included in network expansions strategies (e.g. 'People You May Know' and similar features)

f. Should persuasive design features have been risk assessed as suitably safe they should in any event be disabled. If risks associated with these features cannot be mitigated to acceptable levels then features and functionalities must be designed out. This includes, but is not limited to;

i. Autoplay turned off by default

ii. Infinite scroll turned off by default, and pagination of appropriate length (where used)

iii. Push notifications and alerts turned off by default, and where these can be enabled, sufficient granularity for users to choose which types of notifications they want to see (as opposed to 'all or nothing' approaches)

iv. "Streaks" or engagement rewards disabled

v. Lootboxes or randomised reward features disabled

12 **Prohibition of 'dark patterns' for all users:** Services must not deploy interfaces that manipulate, pressure, or mislead users into making decisions that reduce their safety, privacy, or wellbeing. This prohibition applies to all users, with enhanced protections required where children are likely to access the service.

13 **Content creation:** services should risk assess the tools for the creation of content – this includes but is not limited to bots (including chatbots), bot networks, deepfake or audiovisual manipulation materials and tools, content embedded from other platforms and synthetic features such as gifs, emojis, hashtags.

COMMENTARY

Risk assessment is central to safety by design as it is the mechanism by which providers understand the risks to safety. The United Nations Guiding Principles on Business and Human Rights (UNGP) specifies that companies should have:

> *"a human rights due diligence process to identify, prevent, mitigate and account for how they address their impacts on human rights*[33]*."*

As the Organisation for Economic Co-operation and Development (OECD) guidance notes:

> *"Due diligence is risk-based. The measures that an enterprise takes to conduct due diligence should be commensurate to the severity and likelihood of the adverse impact. When the likelihood and severity of an adverse impact is high, then due diligence should be more extensive.*[34]*"*

Further details about the kinds of steps platforms should be taking can be found in Ofcom's Guidance as well as the Guidance on A Safer Life Online for Women and Girls[35]. Risk assessments should be founded in an intersectional understanding of how harms are directed at, and felt by minoritised groups.

Companies introducing safety by design in their product development for the first time should also evaluate their existing risk management practices and processes to consider any gaps or tensions in those approaches and ensure that there is appropriate governance.

Mitigation strategies should be produced in consultation with independent experts and members/representatives of disproportionately affected groups. This should include:

- That persons who may be affected are consulted by the provider of a regulated service through the entire lifecycle of the service and the functionalities making up the service, including the following stages: design, development, deployment, management, and retirement
- That decisions to be taken about matters which affect the features and functionalities of the service, including content generation, content moderation, content curation, advertising, are not taken without the involvement of persons who may be affected or represent the interests of communities which may be affected
- That governance of the provider of a regulated service includes oversight by persons who may be affected or represent the interests of communities which may be affected
- That due regard is given to tackling how the service and functionalities of the service may discriminate against or harm marginalised groups

EXAMPLE

Somewhere Good: a voice note app, was designed explicitly with the intention of creating a safer online community space for Black and minoritised people, with no content feed, followers, or engagement numbers, limits on posting and setting behavioural norms early on. [36]

Mastodon is a decentralised social network (part of the 'Fediverse'), with separate interoperable servers controlled locally, with moderation decentralised, and no advertising[37]

Good practice indicates that providers should also take into account the entities in their supply chain (for example if buying in third party safety tech, or integrating gen AI based on one of the major models). The OECD notes that there might be practical limitations on the leverage that some entities have over those in their supply chain, and this may change over time. The OECD Guidelines suggest "[a]ppropriate responses with regard to the business relationship may include continuation of the relationship throughout the course of risk mitigation efforts; temporary suspension of the relationship while pursuing ongoing risk mitigation; or, as a last resort, disengagement from a business relationship either after failed attempts at mitigation, or where the enterprise deems mitigation not feasible, or because of the severity of the adverse impact". Allowing third-party tools can

assist users and support safety but they also contribute to a service's risk profile. The ability of a service to be user modified should be considered also.

The hierarchy of control is a significant element of safety by design: it is more effective to make design substitutions or build safety elements in at the beginning that add them retrospectively. Seeking to remove a hazard completely is the most effective approach because there is no risk to manage. Some manipulative design choices should be eliminated entirely; for some persuasive design features substitution for less manipulative versions should be considered. Services should also consider the ease with which ex post safeguards/guardrails can be circumvented, and seek to reduce this possibility to a minimum level.

There has been much concern about anonymous accounts and their role in online abuse. While the code refers to Know Your Customer (KYC) it does not require services to prohibit anonymous accounts - indeed, it should be recognised that anonymity is important for some groups and should be protected as part of ensuring privacy and freedom of expression online[38]. Instead this expects the service to recognise the risks arising from anonymity, and, building on the OSA duties on Category 1 services[39], expect them to take steps to manage that - perhaps by adopting that system even if not a Category 1 service. Moreover, there are risks here relating to multiple accounts, bots and networks of accounts which need to be considered when deciding what is appropriate. In this context, service providers could seek to understand who are the direct and indirect instigators and beneficiaries of such speech, as well as seeking to understand who is operationalising those messages and how (bots, sock puppets networks and false identities etc).

Persuasive design here refers to design choices and functionalities that aim to engage the user, tapping into psychological vulnerabilities all humans have. They utilise the Fogg principles: linking motivation, ability, and triggers at the right moment to drive behaviour. Persuasive design is different from "normal" user interface design, which seeks to make the service easy to use, because of this intentional use of psychological triggers. While we refer here to specific features, there is a close link between persuasive design and dark patterns. For the code, the key point is that persuasive design is about those features that promote repeat or extended use - in non-medical parlance, those features tending towards addiction. We refer to Ofcom's commissioned report on children's financial harm and persuasive design[40]. They identify different sorts of features (a classification which could also apply to dark patterns):

- **Risk-based:** These included mystery rewards, including paid-for rewards, and platforms that visually resembled gambling sites;
- **Dissociative:** Make it harder for users to stay conscious of what and how much they are spending;
- **Misleading**;
- **Impulsive:** Encourage users to make quick, impulsive decisions surrounding spending;
- **Social influence:** Exploit and amplify peer pressures such as sharp differences in default and paid-for cosmetic items – making the paid-for versions feel essential.

These are not necessarily limited to financial harms, however.

When we refer to "dark patterns", sometimes called deceptive design, we refer to decisions relating to the online choice architecture, understood as the contexts in which consumers make decisions and how choices are presented, that nudge or manipulate users to make choices or take actions that those users might not otherwise choose to do. These decisions tend to be harmful to the user but in the service provider's interest. They interfere with the user making informed choices. A key element is the manipulation of the user; there is a distinction between persuasion and manipulation. There is no definition in UK law, though there are some definitions in EU law (see e.g. Recital 67 to the Digital Services Act[41]; EU Commission's fitness check on EU consumer law on digital fairness[42]). The Competition and Markets Authority identified dark patterns as a set of (deliberately) manipulative practices identified by user experience (UX) designers; "sludge" which makes it harder for consumers to cancel or withdraw and "dark nudges" which make it easy for users to take actions not in their best interests (e.g. one-click routes)[43]. Typical examples[44] include fake urgency (for instance fake countdown timers) to pressure user action, disguised advertisement, preselection and emotional manipulation[45], and hyper-nudging[46]. We note also that many "dark patterns" could in any event fall foul of consumer laws on unfair commercial practices, as well as raising questions about compliance with data protection rules[47].

Each service is designed to allow and incentivise a user to create content and communicate with other users in a different way. Features such as metrics or financial incentives based on popularity should be considered in relation to the motivation(s) of the creator. Outrage and content that plays on the biases of users (including racism and misogyny) seemingly drive engagement (as clickbait headlines show), and there is a risk of cycles of ever increasingly outrageous content to drive likes and upvotes. In some cases, the creation and sharing/interacting with highly harmful content can produce engagement and profit for the social media platform[48]. When considering bad actors, services could consider how to disrupt them or at least add friction. Matters that could be considered include: barriers to producing new harmful content; action to prevent the facilitation of harmful contact such as grooming and online exploitation; mechanisms to prevent bad actors directing material at certain users, with the intention to cause harm.

Services should consider how content is created - notably the role of AI content generation tools. Techniques such as those at 6c could be relevant here. Services should also consider their supply chain and what models the tools are built on and the risks that they might import into the service.

PART 3 DEPLOYMENT AND MONITORING

14 TERMS OF SERVICE

Services must clearly define acceptable behaviour and prohibited activity, based on an assessment of risks experienced by users of the services

a. Services should provide a clear framework of use that defines their relationship with users. On the part of users, this should set out acceptable and prohibited behaviour, including harmful or illegal conduct. On the part of the service, it should outline its responsibilities, including any safety measures and default settings in place. Terms of service should also provide information about the company, including its ownership, operational model and how the service is designed, why particular features or flows exist, and how these design choices may influence user behaviour, content sharing, or engagement.

b. Terms of service must expressly prohibit conduct that causes unacceptable harm or breaches the law, including but not limited to harassment, threats, stalking, abusive behaviour, coordinated abuse campaigns and violence against women and girls (VAWG), including digital misogynoir. They must also cover illegal content and harms to children. Terms of service must describe how these risks are continuously monitored and evaluated, including engagement with relevant regulators or authorities to report harms, ensure compliance, and support user safety. These must also be upheld in practice.

c. Terms of service should be able to account for how different groups are targeted in different ways, and prohibit both explicit and implicit forms of violence.

d. Terms of service must recognise and explain that the design and operation of the service (for example recommender systems, reward systems, behaviour-influencing design components, and content-sharing features) can influence how content is created, shared, and amplified, and how users' personal agency may be impacted (e.g., time spent on the service). Services must design and operate these systems in a way that does not incentivise or amplify harm to users.

e. Terms of service must explain how risks are mitigated through regular assessments, safety measures (like limiting reach, adding friction, or redesigning features), and ongoing monitoring to ensure fairness, safety, and minimisation of unacceptable bias.

f. For children, terms of service must be clear, accessible, and tailored to their developmental stage. This means using simple language, breaking content into short, logically organised sections, and providing multiple formats such as visual or interactive explanations where feasible. Terms of service should be easy to navigate, highlight safety information, fully accessible, and include ways for children to understand and give meaningful consent where possible. Services should test terms of service with different age groups and abilities, and regularly update them based on feedback.

g. Services must explain how breaches of terms of service are robustly detected and assessed, whether through automated systems, human review, or user reports, and outline the full range of enforcement actions (e.g. content removal, account restrictions, or suspension) along with indicative timelines for resolution.

h. Reporting mechanisms must be accessible to all users (and non-users where profiles or posts are public-facing), including parents or guardians, and cover sensitive or high-risk situations, such as child bereavement, revenge porn, VAWG, terrorism, or other illegal content. Terms of service should clearly explain how reports are handled, expected timelines, and possible outcomes. Where relevant, terms of service should also describe how the service works with law enforcement or competent authorities.

i. Terms of service should outline procedures for urgent safety situations, like self-harm, suicide, or other immediate threats. This includes rapid escalation, engagement with appropriate support services or authorities, and guidance for affected users.

j. Terms of service must be accessible to all users, taking account of literacy, language, disability, and developmental level. Services should provide transparency about enforcement through aggregate information on actions taken, outcomes, and effectiveness, and test these measures with relevant users to ensure they are understandable and practical.

k. Finally, terms of service must be actively supported by service design. Features, recommendations, and user flows should reinforce rules rather than undermine them. Services must monitor whether design choices - especially those related to engagement or growth - create or amplify risks. In-product prompts, warnings, or friction should support compliance, with appropriate adaptations for children.

COMMENTARY

Terms of service constitute the contract between the service provider and the user. They are important in communicating the service provider's values. As such, they may include community standards (though sometimes Terms of service and Community Standards are used interchangeably) or acceptable use policies, understood as the content and behaviour rules the provider will enforce.

Terms of service should be easily visible before a user signs up to the service, be easy to understand (by the age groups using the service) and be available in languages used by the service's users. This is important as part of transparency, but also to hold service providers and users to account.

Terms of service and Community Guidelines should be kept under review, and revised where appropriate taking into account not just changes in external context but also learning from risk assessments, metrics on effectiveness of mitigation plans and complaints and moderation processes as well as any codes and or guidance from OFCOM.

The Terms of service must explain the steps that regulated services will take if the terms of service are broken by users and be enforced by the online service. Evidence must be kept on individual cases, in line with GDPR requirements, regardless of the final decision. Within the service itself providers must ensure that training and awareness tools are readily available to users on the Terms of service and Community Guidelines to ensure users are aware of permitted content and behaviours on the platforms, and that these tools are kept updated.

15 ROBUST AGE ASSURANCE MECHANISMS

Services must deploy proportionate and highly effective age assurance strategies, depending on the level of risk to children. Age assurance mechanisms must be tested to ensure that they are not implemented in a way that invites children to circumvent the mechanism (e.g. by encouraging repeat tries, offering children under the minimum age for the service the opportunity to try to verify).

a. Services presenting unreasonable risks to certain groups of users (e.g. children) have a responsibility to know their users to ensure that the service they offer is appropriate for the user's age and developmental stage. In many cases, age assurance is a key component of a safety-by-design framework. Where it is feasible and proportionate to implement age assurance, this must be done to enable safe, age-appropriate experiences, not solely to manage or restrict access. However, this is not to say that age assurance is appropriate in all scenarios, nor that it is a 'silver bullet' for online safety.

b. Services must determine whether age assurance is necessary by assessing the specific risks their service poses to children, including the features and functionalities available, the presence of content harmful to children, the likelihood of access by children and the severity of potential harms, alongside any other legal duties. Age assurance measures must be proportionate to the identified risks and designed to prevent children from accessing age-inappropriate features or types of content, while supporting an appropriate user experience. Services should adopt a spectrum of assurance methods tailored to context, risk, and user characteristics.

c. All age assurance methods involve personal data processing and must comply with applicable data protection law. Services must embed privacy preservation by design and default, minimise data collection to that which is strictly necessary, and implement safeguards against misuse, discrimination, and unwanted bias. Systems must not be repurposed for incompatible uses such as marketing or profiling unrelated to safety outcomes, and must be designed and operated in full compliance with UK GDPR, the Age-Appropriate Design Code, ISO/IEC 27566-1 (which provides a general framework for age assurance deployments), and IEEE 2089.1-2024 (which goes into further detail on achieving age assurance that respects children's rights).

d. Services must evaluate the effectiveness of age assurance measures based on outcomes, not solely on the technical selection of methods. This includes measuring whether the system correctly identifies under-age users, correctly identifies adult users, prevents inappropriate access, and reduces identified risks in practice. Performance must be documented publicly against clear criteria, including accuracy, privacy preservation, robustness, reliability, fairness, and user experience, and reviewed regularly.

e. Age assurance systems must be tested to avoid discriminatory outcomes for users with protected characteristics, including race, disability or socio-economic background. Children and relevant user groups should be engaged in testing and validation to ensure systems are understandable, accessible, and effective. Alternative methods must be available for users unable to complete a given assurance method, and transparent challenge and redress process must be maintained. Ideally, multiple alternatives should be offered to allow users unable to complete a given assurance method with a choice of options. For example, for users unable to complete facial age estimation, their only other option should not be to upload a hard ID document where this can be avoided, as this presents outsize risks (e.g. to exclusion, privacy and so on) to groups that are likely already vulnerable (including children).

f. Services must publish information on the performance and impact of age assurance, including rates of correct and incorrect classifications, steps taken to improve accuracy, fairness and accessibility, and actions taken in response to identified performance issues. This transparency must be clear, accessible and updated regularly to enable assessment by users, regulators and other stakeholders.

g. **Age assurance obligations include an ongoing duty to review and update systems, thresholds, and methods in light of emerging technologies, changes in risk profiles, operational experience or new regulatory guidance. Services must implement governance processes to ensure age assurance remains aligned with risk outcomes, user needs, and rights and does not degrade over time.**

h. **Minimum criteria includes:**

   i. Highly effective

   ii. Must offer more than one form of age assurance

   iii. Must be in full compliance with GDPR, AADC and ICO's opinion on age assurance

   iv. Monitored and tested.

   v. Must protect user's rights to privacy, anonymity and access to information

COMMENTARY

Age assurance is a foundational element of the OSA framework, forming part of services' risk management and child protection duties. Beyond merely restricting access, it is central to a broader safety-by-design strategy, ensuring that services are structured, designed, and operated to deliver age-appropriate, safe, and rights-respecting experiences.

Effective age assurance must be risk-based, privacy-preserving, proportionate, and outcomes-focused, measuring whether under-age users are correctly identified, inappropriate access prevented, and risks mitigated in practice. Systems must be transparent, accountable, and regularly reviewed, with performance documented and reported. All methods must comply with data protection law, the Age-Appropriate Design Code, and relevant standards, minimise data collection, prevent misuse or bias, and provide accessible alternatives. Continuous review ensures protections remain effective, aligned with children's needs, and embedded within the service's overall safety-by-design approach.

This guidance builds on 5Rights' work on age assurance[49] and IEEE 2089.1 Standard for Online Age Verification[50] reflecting best practice for risk-based, proportionate, and outcomes-focused approaches.

16 MODERATION

Services must reduce the risks of harmful and illegal activity. Systems that promote or recommend content or allow for its curation (e.g. via tagging systems) must not systematically amplify harm. Services must use a proportionate mix of both automated and human moderation systems when tackling illegal and harmful activity. This should include continuous post-market monitoring and red-teaming as relevant.

a. Services must conduct periodic reviews of the efficacy of their enforcement of their terms of service, including through engagement with affected users, with particular regard to ways in which enforcement may be disproportionately affecting marginalised groups. They must take action to remediate systemic over-enforcement against the content or activity of members of marginalised groups or systemic under-enforcement of certain kinds of harmful content or activity. Enforcement systems should also be audited and safety-testing of systems must include how context-sensitive they are.

b. Services must actively manage risks and protect users from harm throughout deployment. Moderation systems should cover both (a) user-generated content and activity, including posts, comments, images, videos, interactions, and (b) system-driven features such as recommendations or notifications. Moderation systems should be able to find content harmful to children or illegal content and activity, take action to deal with it, and report it where necessary.

c. For children, additional safeguards are required. Moderation must take account of developmental vulnerabilities such as increased credulity, ensuring that content and interactions do not exploit or expose children to preventable harms. This includes careful oversight of algorithmic amplification and recommendation, age-appropriate content filters in line with the evolving capacity of the child, strong default safety settings, and reporting and redress mechanisms that are accessible and user-friendly for children.

d. Moderation should be delivered through a mix of human review, automated systems, and hybrid approaches, with each method designed to ensure accuracy, fairness, and timeliness. Automated tools can potentially detect patterns, keywords, or known harmful content at scale, while human moderators provide context-sensitive judgment in complex or borderline cases. Hybrid systems combine the speed of automation with the nuance of human oversight, while ensuring compliance with human rights law, including protection of freedom of expression, non-discrimination and the United Nations Convention on the Rights of the Child. Measures should be taken to minimize the negative impacts of automation bias and human unconscious bias in moderation. Solely automated moderation (with no human oversight) must be avoided.

e. Moderation functions must be suitably resourced to carry out their role. This includes (but is not limited to) resources such as available staff, department budget and training. Moderation teams also need sufficient authority and independence to carry out their roles effectively.

f. Moderation teams must be adequately protected and compensated, including provision of psychological support. This should apply to outsourced moderation as well as in-house.

g. Moderation systems must be designed to anticipate potential risks in addition to responding to known harms. Services should take account of testing carried out at the design stage, including red teaming, and continue to assess for potential harms during deployment through ongoing testing, monitoring, and user feedback. Services must establish feedback loops from user reports, internal data analytics and external audits.

h. Moderation should be a board-level responsibility. Senior managers must receive regular reports on the effectiveness of moderation systems, emerging risks, and patterns of harm. Services should conduct audits, working with members of affected groups and organisations which represent them, to identify any disproportionate impact on marginalised groups and take corrective action as necessary. Key performance indicators for moderation should align with broader service design and safety objectives, rather than focusing solely on content removal, ensuring governance decisions actively support the service's safety-by-design strategy.

i. Services should provide transparency through public dashboards reporting aggregate moderation outcomes, algorithmic impact assessments, and policy updates. User-facing transparency must also be maintained, clearly explaining moderation policies, appeal mechanisms, and the rationale for content removal decisions. Appeal mechanisms should be age-appropriate, accessible, and easy to use. It should always be possible to appeal automated moderation decisions.

COMMENTARY

While Ofcom's Codes set out the minimum expectations for moderation, primarily focused on detection, reporting and enforcement of illegal content and content harmful to children, a safety-by-design approach goes further. International technical standards, such as the IEEE Standard 7010 on Age-Appropriate Design and Safety of Online Services provides guidance for anticipating harms, embedding oversight into product development, and measuring outcomes rather than processes alone.

In practice, this means that moderation is integrated throughout the lifecycle of the service, covering both content and activity. Hybrid systems, combining human judgment and automated tools, must be calibrated for accuracy, fairness and inclusivity, with protections for freedom of expression, non-discrimination and alignment with the United Nations Convention on the Rights of the Child.

Moderation should be integrated into the design of the system and the user experience, embedding safety considerations directly into product features. For example, comments can be reviewed before they are made visible, limits can be

placed on messaging with strangers, and recommendation systems can be designed to avoid amplifying harmful or extreme content. Default safe settings should be applied so that users begin in a safer mode unless they actively choose to change them.

Safety-by-design moderation also places governance and accountability at board level, with reporting on effectiveness, emerging risks, and impacts on marginalised groups. Key performance indicators must be tied to outcomes, not just content removal metrics. Transparency is enhanced beyond the regulatory minimums in Ofcom's codes and would include public dashboards, algorithmic impact assessments and clear explanations of policies, enforcement, and appeals.

This principle is underpinned by IEEE Standard 2089 on Age-Appropriate Design[51], which provides technical guidance on designing services that anticipate and mitigate risks to users, particularly children, while embedding protections into the lifecycle of the product or feature.

EXAMPLE

Chelsea Peterson-Salahuddin describes how hate speech content moderation could be done with reparation in mind, which focuses on "redress, not removal" - redress which could include apologies from the poster of hateful content, affirmative actions, and/or wider community development schemes involving marginalised users to identify further ways that harm could be mitigated and repaired.[52]

Danielle Cole similarly describes a Black feminist model of social media, in which a platform would facilitate time-outs and 'calling in' of users who have shared hateful content, and accountability for behaviour would be shared by the online community, focusing on repair rather than punishment.[53]

17 USER TOOLS

Users must be provided with effective tools to control their experience and protect themselves. These include tools to report, content, features and functionality, options to reset or select content curation algorithms, blocking and reporting of other users, tagging controls, and controls over reply.

a. Services must provide meaningful, accessible user tools that allow individuals to shape their experience of the service and manage exposure to categories of lawful content that could still be harmful or distressing, such as violent material or offensive language. These tools should empower users by giving them control over content, interactions, notifications, and other features that affect safety and exposure to risk. User tools should be designed to anticipate potential harms, provide proactive guidance, and integrate seamlessly with other safety mechanisms.

b. Tools must reduce harm without unnecessarily restricting other content, allowing users to tailor protections to their needs. They should complement systemic safeguards, including governance, algorithmic controls, and risk assessments, offering a layered and coherent approach to safety. Services must embed privacy and data protection by design and default, ensuring compliance with UK GDPR, the Age-Appropriate Design Code, and any relevant technical standards. Data collection must be minimised, with safeguards in place to prevent misuse, discrimination, or profiling.

c. User tools must be transparent, understandable, and effective. Services should monitor usage, assess effectiveness based on outcomes, and publish reporting to enable regulatory oversight. Tools must be updated continuously based on diverse user feedback and other available evidence to address emerging risks, user feedback, technological change, and regulatory guidance. Governance processes should ensure tools remain effective and do not degrade over time. Where tools enable users (or non-users) to contact the organisation directly, there must be a named person or person(s) accountable for responding.

d. Tools must be tested with a range of relevant user groups, including children, to prevent discriminatory outcomes. Engagement with users, including children, must be participatory and respectful, not tokenistic or leading. Alternative, accessible methods should be provided for users unable to use standard mechanisms, and transparent processes must be maintained for challenge and correction. Tools should also offer guidance and real-time feedback to help users understand why protections are triggered and how to adjust settings safely.

e. User tools must work coherently with other safety features, including risk assessments, moderation, and age assurance. Services should clearly communicate how these tools contribute to overall safety and support personal agency.

f. For children, services must provide age-appropriate tools that embed safety and support throughout the user experience. This includes default safety settings and friction mechanisms tailored to the child's age and developmental stage; integration with age assurance; moderation, and reporting systems to prevent exposure to age-inappropriate content; and accessible alternatives for children with disabilities. Services should provide clear crisis support pathways for children and caregivers, including signposting to trained professionals or helplines in cases of abuse, self-harm, or other urgent risks, alongside interactive guidance and education to promote digital literacy and safe online engagement. Children's tools must be continuously evaluated and iterated based on user testing, risk assessments, and emerging threats, ensuring protections remain effective, equitable, and aligned with rights-respecting design principles.

COMMENTARY

While Ofcom's codes require some services to provide reporting mechanisms and baseline safety settings, particularly for children, they are primarily focused on compliance and reactive safeguards. A safety-by-design approach goes further, embedding user tools that empower agency and self-management, supporting diverse needs and vulnerabilities. It also links governance and design decisions to measurable safety outcomes. This approach aligns with technical standards and best practice on age-appropriate design, including IEEE Standard 2089[54] and positions transparency and user empowerment at the core of product strategy rather than as an adjunct to enforcement.

EXAMPLE

Bluesky has algorithmic feed options that users can choose from, and developers can build new feeds.[55]

18 TRANSPARENCY

a. Services must have a comprehensive and proactive strategy for providing stakeholders with clear and accessible information about the design, operation and remaining risk present on the service. This must include the risks identified through ongoing risk assessments, including those arising from features, functionalities, and business models, and details of the specific mitigations implemented, and how the effectiveness of those mitigations are measured. It must also include clearly assigned points of accountability to the public and to regulators.

b. Services should publish regular, clear transparency reports that allow regulators, researchers, and the public to understand how the service manages risk. These reports should show how much harmful or illegal content is detected and acted on, what types of enforcement actions are taken and how quickly after being made aware of the harm, service response to user reporting and how quickly after a report is received, and how any automated safety tools are working. They should also report on effectiveness, including accuracy, errors, and whether risks are being reduced over time. Additionally, reports should cover user complaints, reports, and appeals (e.g. how many there are, how they are resolved, and how long responses take). The metrics and measures included in this reporting must be appropriate to deliver the full picture of safety, and must not be presented in misleading or manipulative ways.

c. Services should organise information into clear, meaningful categories, such as the type of risk, the feature involved, and the affected user group. They should highlight any systemic risks, emerging harms, and the steps taken in

response. Transparency measures must be easy to understand and accessible to different audiences (e.g. through tailored summaries) and they should allow independent scrutiny, including by accredited researchers. Reporting should be part of an ongoing governance cycle, helping services continually assess risk, improve design, and refine safety measures.

d. Information shared in transparency reporting should also include how different groups, in particular marginalised groups, are being affected by risks on the service and how effective mitigations are in reducing these risks.

e. When children are likely to access a service, transparency reporting must specifically highlight risks to children and caregivers and how effectively those risks are being managed. This should include age-specific data, information on harms affecting children, how age assurance and age-appropriate design measures are working, and evidence that children's rights, safety, and developmental needs are being respected. Information aimed at children and caregivers should be presented in an age-appropriate, clear, and accessible way, so that it can be easily understood and used for informed oversight.

COMMENTARY

Transparency is a central element of a safety-by-design approach, supporting accountability, oversight, and empowering users to make informed choices. Under Ofcom's Codes, services are required to provide information on rules, reporting mechanisms, moderation processes, and to submit formal reports and regulatory disclosures of risk assessments and safety measures. These requirements, however, are limited in scope and some apply only to specific services.

A safety-by-design approach goes further, requiring services not only to describe the systems they have in place, but to show how effective those systems are in reducing harm. Transparency should cover how risks are identified, how design choices shape user experience, and whether mitigation measures are achieving their intended outcomes.

To be meaningful, transparency must be accessible and user-focused, particularly for children, with information presented in a way that is understandable and actionable. Transparency must serve as a tool for continuous improvement and safer service design, rather than a box-ticking exercise. Furthermore, it must not be undermined by internal conflicts of interest, requiring a culture of openness and non-retaliation when personnel raise problems.

This guidance is underpinned by the UK Age-Appropriate Design Code (AADC)[56], which sets statutory requirements for services to provide clear, accessible, and age-appropriate information to children. It also draws on 5Rights' Tick to Agree[57]

framework ensuring terms of service are understandable, user-centred, and support informed decision-making

19 REDRESS MECHANISMS

a. Services must offer clear and accessible ways for users to seek help and remedies when they experience harm (including discrimination and bias through other safety measures on the service). This includes straightforward processes for removing content, taking action on accounts, and preventing repeat incidents, as well as challenging decisions that may unfairly restrict or demote user's content despite it being non-violative, or restrict access or features due to a user's age, disability, or other protected characteristics. Redress should be easy to understand, transparent and backed up with clear information on what users can expect, including timelines and appeal options. Remedies should be fair, consistent, and tailored to the seriousness and context of the harm, addressing both individual cases and underlying system-wide issues. There must be a person or person(s) within the organisation responsible for redress procedures – reviews cannot be solely automated.

b. Redress should not exist in isolation. Insights from complaints, reports, and appeals should feed back into the design and operation of the service, helping to prevent future harm and improve safety measures across the platform.

c. For children, redress must be age-appropriate, prioritising protection, support, and learning rather than purely punitive action. Services should respond in ways that match children's developmental stage, helping them understand harmful behaviour, the consequences, and how to stay safe online. Systems should provide clear options for children to report harms or challenge decisions, with any interventions being proportionate, supportive, and aligned with their capacity to understand. Signposting to support and helpline services must be made available for children.

d. Services should also ensure that children and their caregivers have access to appropriate support in sensitive situations. Enforcement actions should be proportionate and transparent, aiming wherever possible to guide safer behaviour and encourage positive engagement rather than simply excluding the child from the service.

e. Services should provide signposting and funding to external services which support victim-survivors of online harms, in particular by-and-for services supporting Black and other minoritised groups.

COMMENTARY

Redress mechanisms are a critical element of a safety-by-design approach,

extending beyond the procedural focus of Ofcom's online safety codes. While the Act primarily sets out how services should respond to illegal content, a robust redress framework ensures that users can obtain timely, proportionate, and meaningful remedies, and that insights from complaints inform ongoing improvements to system design and safety measures.

For children, redress must be tailored to developmental needs, prioritising protection, guidance, and learning rather than solely punitive action. Effective mechanisms support children in understanding harmful behaviour, recognising consequences, and engaging safely online. This approach is underpinned by the IEEE 2089 on age-appropriate design[58], ensuring that redress processes are designed in alignment with children's rights, developmental capacities, and the broader objectives of a rights-respecting, safety-by-design framework.

EXAMPLE

Platforms could recognise their complicity in harms to users from tech-facilitated abuse, such as non-consensual intimate image sharing on tech platforms which is not adequately remediated through content takedown. This could look like platforms apologising directly to victim-survivors affected on their services, as well as platforms contributing financially to access to therapies, compensation, healing activities and support for victim-survivors.[59]

Bumble, a dating app, partnered with Chayn, a nonprofit which centres survivors and uses trauma-informed design principles[60], to provide online support and in some cases therapy sessions to victim-survivors of sexual violence or abuse.[61]

PART 4 RETIREMENT AND DECOMMISSIONING

20 RETIREMENT AND DECOMMISSIONING

a. Services must ensure that the removal or shutdown of features or services does not create additional risks to users. This includes:

- **i.** Users must receive reasonable notice before services or features are withdrawn.
- **ii.** Personal data must be securely deleted in accordance with the UK GDPR.
- **iii.** Where applicable, users must be able to exercise their rights to data portability.
- **iv.** Safety transition planning (e.g. where it may lead to loss of moderation safeguards or migration of harmful communities to unregulated spaces.)
- **v.** Archiving and evidence preservation

b. Services must manage the retirement, withdrawal, or significant modification of features, systems, or the service itself in a way that prioritises user safety and maintains existing protections. Services should determine appropriate triggers for decommissioning procedures (e.g. if risks remain at an unacceptable level even after all mitigations are applied) in advance, preferably during the initial design phase.

c. Before decommissioning any feature or system, services should conduct a forward-looking risk assessment to identify how its removal may create new risks, weaken existing safeguards, or affect vulnerable users, particularly children. This includes reviewing dependencies on safety features, moderation tools, reporting mechanisms and user controls that could be impacted during transition. This could also include reviewing access to certain parts of the service which may be blocked to certain groups of users (e.g. children).

d. Clear transition plans must be put in place to maintain or enhance protections, including the continued operation of moderation, reporting, and redress mechanisms throughout the decommissioning process. Users should be given timely, clear, and accessible information about changes, including how their data, content and safety settings will be affected, and any actions they may need to take.

e. User data must be handled responsibly at end-of-life, in line with data protection law. This includes minimising unnecessary retention, pseudonymizing or anonymizing as appropriate, securely deleting data where appropriate, and ensuring that data collected for safety purposes is not repurposed in ways that undermine user rights or expectations.

f. Decommissioning processes must also include mechanisms for user support and redress, especially where changes could expose users to new risks or disrupt protections. Services should actively monitor the impact of

decommissioning decisions and take corrective action if unintended harms arise.

g. For children, additional safeguards are required. Services must provide age-appropriate communication, maintain protections, and offer support to children and caregivers during transitions. Decommissioning must not reduce safety or increase exposure to harm for children.

h. Finally, all decommissioning decisions should be documented, reviewed, and incorporated into ongoing governance and safety processes. Lessons learned must inform future design, deployment, and retirement of systems, ensuring continuous improvement in protecting users throughout the service lifecycle.

COMMENTARY

Decommissioning is a crucial but often overlooked stage in the product or service lifecycle. It implicates activities such as communication with stakeholders, proper data destruction and/or disposal, and dependency management. For children, all communications and support relating to decommissioning and service shutdown must be clearly presented in a manner they can understand.

Organisations need advanced knowledge of when they should retire their offerings, in order to sufficiently plan ahead. For instance, on trigger for stopping or decommissioning a project is when residual risk is found to remain at unacceptable levels even after all possible mitigations have been applied.

ENDNOTES

1 https://www.theguardian.com/media/2026/mar/25/jury-verdict-us-first-social-media-addiction-trial-meta-youtube

2 https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/research-statistics-and-data/online-services-research/persuasive-design-features-and-potential-child-financial-harms-report.pdf?v=399304

3 https://mollyrosefoundation.org/wp-content/uploads/2025/08/proof3_PervasivebyDesign.pdf

4 https://riskybydesign.5rightsfoundation.com/recommendation-systems

5 https://www.onlinesafetyact.net/documents/1189/OSA_Network__a_10-point_plan_for_Government.pdf

6 https://mollyrosefoundation.org/wp-content/uploads/2026/02/MRF-Roadmap_Briefing-Feb-2026.pdf

7 https://www.internetmatters.org/hub/research/childrens-wellbeing-in-a-digital-world-index-report-2026/

8 Section 1(3) OSA 2023: https://www.legislation.gov.uk/ukpga/2023/50/section/1

9 https://www.ofcom.org.uk/online-safety/protecting-children/remarks-by-melanie-dawes-ofcom-chief executive-at-the-nspcc

10 https://digital-strategy.ec.europa.eu/en/policies/digital-services-act

11 https://www.esafety.gov.au/industry/safety-by-design

12 https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf

13 https://docs.un.org/en/A/HRC/61/45

14 https://www.iso.org/standard/63787.html

15 https://designjustice.org/principles-overview

16 https://designjustice.mitpress.mit.edu/

17 https://www.wired.com/story/drag-queens-vs-far-right-toxic-tweets/

18 https://www.theguardian.com/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive

19 https://designjustice.mitpress.mit.edu/

20 https://direct.mit.edu/books/oa-monograph/4605/Design-JusticeCommunity-Led-Practices-to-Build-the

21 https://designjustice.org/

22 https://www.legislation.gov.uk/ukpga/2023/50?view=plain

23 https://www.gov.uk/government/publications/statement-of-strategic-priorities-for-online-safety/final-statement-of-strategic-priorities-for-online-safety#safety-by-design

24 https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/consultation-on-draft-guidance-a-safer-life-online-for-women-and-girls/statement-docs/guidance-a-safer-life-online-for-women-and-girls.pdf?v=408215

25 https://www.onlinesafetyact.net/analysis/safety-by-design/

26 https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/register-of-risks.pdf?v=390983

27 https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984; https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/statement-protecting-children-from-harms-online/main-document/childrens-risk-assessment-guidance-and-childrens-risk-profiles.pdf?v=396653

28 https://standards.ieee.org/ieee/2089/7633/

29 https://www.iso.org/standard/27001

30 FCA (2026) Senior Managers and Certification Regime

31 https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/illegal-harms/risk-assessment-guidance-and-risk-profiles.pdf?v=390984

32 Where there are conflicting interests or rights at play, the best interests of the child (as strictly defined in the UN Convention on the Rights of the Child) must be prioritised. Where

there are trade-offs or conflicts implicating multiple rights children hold, proper balancing exercises are to be conducted in line with UNCRC General Comment 14.

33 https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf

34 https://www.oecd.org/en/publications/oecd-guidelines-for-multinational-enterprises-on-responsible-business-conduct_81f92357-en.html

35 https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/consultation-on-draft-guidance-a-safer-life-online-for-women-and-girls/statement-docs/statement-guidance-on-a-safer-life-online-for-women-and-girls.pdf?v=408227

36 https://www.businessinsider.com/somewhere-good-social-media-app-twitter-alternative-bipoc-2022-11

37 https://joinmastodon.org/

38 https://www.ohchr.org/en/stories/2015/06/human-rights-encryption-and-anonymity-digital-age

39 OSA Section 64: https://www.legislation.gov.uk/ukpga/2023/50/part/4/chapter/1/enacted

40 https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/research-statistics-and-data/online-services-research/persuasive-design-features-and-potential-child-financial-harms-report.pdf?v=399304

41 https://www.cms-digitallaws.com/en/dsa/recital-67/

42 https://commission.europa.eu/law/law-topic/consumer-protection-law/review-eu-consumer-law_en

43 https://www.drcf.org.uk/siteassets/drcf/pdf-files/harmful-design-in-digital-markets-ico-cma-joint-position-paper.pdf?v=380506

44 https://mpo.gov.cz/en/consumer-protection/eu-and-the-consumer/behavioural-study-of-the-european-commission-on-dark-patterns-in-the-digital-environment--268126/

45 https://www.gov.uk/government/publications/online-choice-architecture-how-digital-design-can-harm-competition-and-consumers/evidence-review-of-online-choice-architecture-and-consumer-and-competition-harm#oca-practices

46 https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.329

47 https://policyreview.info/articles/news/emergence-of-dark-patterns-as-a-legal-concept

48 https://www.science.org/doi/10.1126/sciadv.abe5641

49 https://5rightsfoundation.com/wp-content/uploads/2026/02/5rights_AgeAssurance_FINAL.pdf

50 https://standards.ieee.org/ieee/2089.1/10700/

51 https://5rightsfoundation.com/wp-content/uploads/2024/09/2089-2021-with-disclaimer.pdf

52 https://journals.sagepub.com/doi/full/10.1177/20539517241245333

53 https://www.tandfonline.com/doi/full/10.1080/14680777.2021.1954970

54 https://5rightsfoundation.com/wp-content/uploads/2024/09/2089-2021-with-disclaimer.pdf

55 https://bsky.social/about/blog/7-27-2023-custom-feeds

56 https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/

57 https://5rightsfoundation.com/resource/tick-to-agree-age-appropriate-presentation-of-published-terms/

58 https://5rightsfoundation.com/wp-content/uploads/2024/09/2089-2021-with-disclaimer.pdf

59 https://static1.squarespace.com/static/67c75573c4cb1959a4a30215/t/68c921aef040d208d19fa5c9/1758011822475/Beyond+Content+Takedown_FINAL+WEB.pdf

60 https://www.chayn.co/about

61 https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/263963-categorisation-research-and-advice/responses/bumble?v=293131

SAFETY BY DESIGN CODE OF PRACTICE

2026

Co-produced with Online Safety Act Network, 5Rights Foundation, End Violence Against Women Coalition, Glitch, Internet Watch Foundation, Refuge, NSPCC, Molly Rose Foundation, FlippGen, and Professor Lorna Woods OBE