ENFORCING THE ONLINE SAFETY ACT FOR CHILDREN

Ambitions for the Children's Safety Code of Practice

April 2024

Photography: Amina Filkins

5RIGHTS FOUNDATION

NSPCC
BARNARDO'S
CCDH
nwg
Childnet
NATIONAL CHILDREN'S BUREAU
ANTI-BULLYING ALLIANCE
BRECK FOUNDATION
LUCY FAITHFULL FOUNDATION
ecpat UK Every Child Protected Against Trafficking
ONLINE SAFETY ACT NETWORK
MCF Marie Collins Foundation
CEASE
Reset.
END VIOLENCE AGAINST WOMEN
kidscape Help With Bullying
The Children's Media Foundation
UK Safer Internet Centre
IWF Internet Watch Foundation
Clean up the internet
BEREAVED FAMILIES FOR ONLINE SAFETY

Introduction

The 2023 UK Online Safety Act is a watershed moment for children. Tech companies will be required to make their services safe by design and to offer children a higher standard of protection than adults.[1]

Next month, Ofcom will publish its draft guidance and codes of practice on the child-specific duties in the Act. It will not be sufficient to codify current industry practice. Ofcom must be ambitious in what it requires from tech companies whose services are used by children in the UK and focus on outcomes rather than prescriptive rules or 'tick-box' processes.

This means requiring regulated service providers to:

1. Give high standards of protection to children using high-risk services, irrespective of the size of the service;
2. Prioritise children's safety in product design and development;
3. Take a comprehensive approach to risk mitigation that considers age-appropriate access to content, features and functionalities, safety and privacy settings, user reporting, media literacy, and the advice of external experts and children themselves.
4. Give safety teams sufficient resources and autonomy to prioritise children's best interests, even when these conflict with commercial interests; and
5. Consider the impact of their business model on safety and ensure governance and accountability checks and balances are strong.

This document sets out the recommendations of the Children's Coalition for the Online Safety Act on what regulated services must do to meet their child safety duties. It draws on the deep expertise and experience of our organisations. It is underpinned by the United Nations Convention on the Rights of the Child[2] and its General comment No. 25.[3] Our recommendations honour the intention of Parliament. They also reflect the lived experiences and voices of children and anticipate that children in different age groups or with certain characteristics and experiences may have different or heightened needs.

[1] s.1, Online Safety Act 2023

[2] United Nations (1989) United Nations Convention on the Rights of the Child

[3] United Nations Committee on the Rights of the Child (2021) General comment No. 25 on children's rights in relation to the digital environment

As a community of experts and as individual organisations, we look forward to our continued engagement with Ofcom to ensure children are safe and able to thrive online.

5Rights Foundation
Barnardo's
NSPCC
Online Safety Act Network
Center for Countering Digital Hate
Breck Foundation
National Children's Bureau
Anti-Bullying Alliance
Clean Up the Internet
Children's Media Foundation
Bereaved Families for Online Safety
NWG Network
Marie Collins Foundation
End Violence Against Women
UK Safer Internet Centre
Reset
CEASE
Kidscape
Lucy Faithfull Foundation
Internet Watch Foundation
ECPAT UK

Safety by design and a higher standard of protection for children

The design of services plays a fundamental role in directing children towards and away from risk. Safety by design[4] requires services to embed safety into all stages of its product design process so that risks are identified and mitigated *before* harm occurs. The requirement to ensure regulated services are safe and age-appropriate by design is an overarching duty set out in Section 1 of Online Safety Act, and it intersects with all other compliance duties and measures.

In its child safety duties codes and guidance, Ofcom must require that:

- The decision to launch a new product, feature or functionality is not taken until a full assessment of risk has been undertaken and senior leadership is satisfied that those risks have been removed or effectively mitigated.

- Service providers consider all aspects of the design of their service that pose particular risk. This includes reviewing the existing design - including the features, functionalities, and business model of the existing service, in addition to any changes going forward.[5]

- Product testing and external consultation with child safety experts is carried out during the design and development of functionalities, algorithms, and other features to identify whether those features or functionalities are likely to contribute to the risk of harm arising from content on the service.

[4] UK Government (2021) Principles of safer online platform design

[5] Features and functionalities that pose particular risk include recommender systems, livestreaming, end to end encryption, pseudonymity and anonymity, live chat (audio and video), messaging (group and private), social features and connections, ephemeral content, persuasive design, gifting, metaverse virtual spaces, and geolocation tracking and sharing.

Photography: Annie Spratt

“Services should assume that young people are going to be on it because young people are all over the internet: the internet is our oyster.” – 5Rights youth adviser

“If you are in a bad mood or feeling low you can be more attracted to a depressing post on social media.”[6] – Child who spoke to Barnardo’s

“Services should actually think about the negative effects their features have, even if they don’t mean for those effects to happen. They shouldn’t just push out new features and then wait and see what happens.” – 5Rights youth adviser

[6] Barnardo’s (2021) Left to their own devices: Young people, social media and mental health

1. Child access assessments (CAAs)

Regulated service providers will need to assess whether children are likely to access the service or parts of the service. If they can, services must comply with the child safety duties in the Act. Ofcom is required to publish separate guidance on Child Access Assessments (CAA).

When drafting its CAA guidance, Ofcom must ensure that:

- The Act applies to the services children are using in reality, and not only to services targeted exclusively at them. This is what was intended by Parliament when it used the language of 'likely to be accessed by children.'[7]

- A 'significant number of child users' must be clearly defined by Ofcom as being "*more than a de minimis or insignificant number of children using the service*." This is in line with existing regulation under the Information Commissioner's Office's Age-Appropriate Design Code.[8]

- Smaller but high-risk services are fully in scope of the child safety duties. Access to small high-risk services, such as suicide forums, can have catastrophic consequences for children. These must not be exempted and must be subject to the strictest measures in the regime, as intended during the passage of the Act when the Government stated that there would be "no exemptions for small companies."[9]

[7] Lord Parkinson of Whitley Bay, 25th April 2023, Online Safety Bill, Committee (2nd Day), col. 1137

[8] Information Commissioner's Office (2021) Introduction to the Children's code

[9] Lord Parkinson of Whitley Bay, 19th July 2023, Online Safety Bill, Report (5th day), col. 2345

2. Age assurance

Highly effective age assurance is important not just for protecting children from content harmful to them, but for activating all child safety measures, including default privacy settings which can protect children from grooming. The use of age assurance and age verification must consider children's fundamental rights as set out in the UN Convention on the Rights of the Child[10] and General Comment 25.[11] It must be used to create positive experiences for children online and must not be used to keep them out of services which, with robust risk mitigation in place, could be age appropriate.

In its child safety duties codes and guidance, Ofcom must require that:

- Age verification is highly effective in preventing access to primary priority content that is harmful to children. Ofcom must provide clear guidance on what 'highly effective' means and compliance must be based on outcomes and efficacy.
- Where age assurance is considered appropriate, it must be proportionate, effective, open to challenge, accessible and inclusive.[12]
- Any user's data used to assess age is gathered and processed with strict adherence to UK data law including the Age Appropriate Design.[13] It must not be used for any other purpose, and services must delete the data once it has served its purpose.

[10] United Nations Convention on the Rights of the Child

[11] UNCRC General comment No. 25

[12] 5Rights Foundation (2021) But how do they know it is a child? pp. 48-53.

[13] Ibid. p48

3. Children's risk assessments (CRAs)

A comprehensive children's risk assessment, which anticipates and mitigates risk of harm, is fundamental to a child-centred and rights-respecting online safety regime. Children's risk assessments must be comprehensive, anticipate and identify risks *before* they emerge and adopt mitigation measures that meet safety by design principles. At a minimum, they must follow and document these four steps:

1. **Understand** online harm to children, including cumulative harms and harms that stem from features and functionalities, used alone or in combination

2. **Assess** the risk of harm to children of different ages and developmental stages considering each risk individually and collectively. Services must also consider heightened risk for vulnerable children.
3. **Decide** and implement prevention and protection measures. Record the criteria used and testing undertaken at all stages of the decision-making process. Adopt mitigation measures that meet safety by design principles.

4. **Review** the assessment of risk regularly, report any changes and update measures accordingly.

In its child safety duties codes and guidance, Ofcom must require that:

- Service providers consider the impact of their business model in creating or reducing risk.

- Children and independent child safety experts are consulted to help companies anticipate and respond to risk.

- Service providers monitor and measure the effectiveness of its risk mitigation strategy.

- Services implement the highest existing technical standards, such as IEEE Standard 2089 for an Age-Appropriate Digital Services Framework.[14]

[14] IEEE Std. 2089 (2021) IEEE Standard for an Age Appropriate Digital Services Framework

“I was playing an online game and these girls kept bullying me because I said I was Indian. They kept saying that Indians are gross and wouldn’t leave me alone. I would leave a game and join another server but they followed me and carried on being racially abusive until I left all together. I have reported them but I doubt it will work.” – Girl aged 12, Childline

“If services are going to have things that are not safe for children or teens, then the responsibility is on them to know which of their users are children or teenagers and which are adults. That way they can make sure that younger children have a different kind of experience, with more support, and older teens can maybe have more options – but we don’t want all the adult things that they push on us.” – 5Rights youth adviser

“Social media can go from helpful to unhelpful to harmful. There is a fine line between what is beneficial and detrimental for mental health. The crossover can be easy.”[15] – Child who spoke to Barnardo’s

[15] Left to their own devices: Young people, social media and mental health

Photography: Andrey K.

4. Content and conduct that is harmful to children

We welcome the broad definition of 'content' within the Act, which includes harm caused by other people's behaviour or conduct or contact with other users.

In its child safety duties codes and guidance, Ofcom must establish clear and robust requirements concerning:

- **Types of harmful content:** While some specific harms are set out in the Act, regulated service providers must take a proactive approach to identifying other types of content that might be harmful to children or children in certain age groups – for example themes of sex which do not meet the threshold of pornography, sexual violence, nudity, horror, and bad language. Services should have regard to the age-appropriateness of content to children in certain age groups and seek to strike a satisfactory balance between what younger children should be protected from and what older children may have a right to access.

- **Heightened vulnerabilities:** Regulated service providers must consider how children with certain characteristics or experiences might be more vulnerable to certain kinds of content than others. For example, victims of sexual abuse might be more vulnerable, and some children will be more likely to be targeted in relation to their gender and other intersecting inequalities.

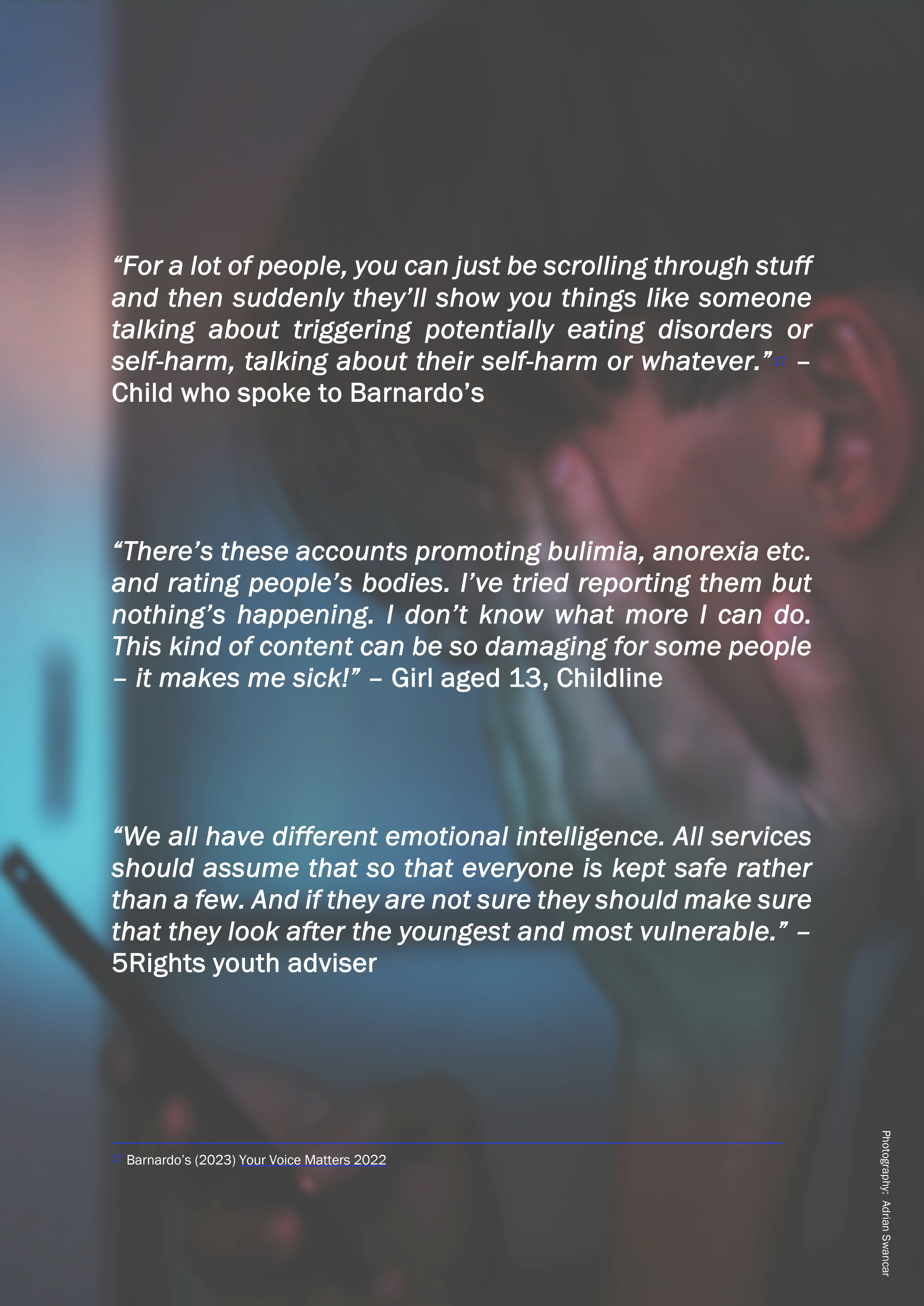
- **Features and functionalities:** Systems and processes to mitigate risk of harm must operate across all features and functionalities of a service including content feeds, account profiles, group pages, comments, livestreaming, and messaging. Where a risk cannot be sufficiently mitigated, services should prevent all children or children in certain age groups from accessing the relevant features or functionalities. Service providers must regularly engage with children and child advocates to understand how children are using their services, and any challenges they are experiencing. Services must never make children choose between functionality and privacy and safety.

- **Recommender systems:** Children are not only vulnerable to risk in content feeds, but to many different algorithmic models. Regulated service risk assessments must identify and assess all algorithmic models deployed across a service and their role in determining what content children interact with on a platform. This must include models that order, recommend and rank videos, images, comments, search terms and results, accounts to follow and friend, comments, tournaments, events, livestreams etc. It also must account for how the child's profile is shared or recommended to other users.

- **Extended use and persuasive design:** Regulated service providers must audit and then remove or restrict design features which undermine a child's agency, exploit their evolving developmental capacities, or have the effect of keeping children on a regulated service for extended periods when it is not in their best interests. This includes persuasive design features (also known as dark patterns) such as auto-play functions, push notifications, endless scroll, random-reward features, popularity metrics, incentives to produce and share content, and techniques to apply time pressure or build anticipation. Regulated service providers must identify opportunities to meaningfully offer children agency about how much time they spend on a service.

- **Cumulative harm:** Regulated service providers must understand how volume and concentration of content can drive risk of harm and take a 'by design' approach to mitigating cumulative harm risks. Cumulative risk of harm includes the combined use of features and functionalities within a service; recommender systems that determine the volume and concentration of content ('dosage'); and persuasive design strategies that extend the amount of time a child spends on a service. Regulated service providers must consider whether the threshold for primary priority, priority and non-designated harmful content has been reached through cumulative exposure as well as individual pieces of content. Moderation systems must be effective in monitoring risk and mitigating residual cumulative harm.[16]

[16] Ofcom (2023) Media literacy by design: Best practice principles for on-platform interventions to promote media literacy



Photography: Adrian Swancar

"For a lot of people, you can just be scrolling through stuff and then suddenly they'll show you things like someone talking about triggering potentially eating disorders or self-harm, talking about their self-harm or whatever."[17] – Child who spoke to Barnardo's

"There's these accounts promoting bulimia, anorexia etc. and rating people's bodies. I've tried reporting them but nothing's happening. I don't know what more I can do. This kind of content can be so damaging for some people – it makes me sick!" – Girl aged 13, Childline

"We all have different emotional intelligence. All services should assume that so that everyone is kept safe rather than a few. And if they are not sure they should make sure that they look after the youngest and most vulnerable." – 5Rights youth adviser

[17] Barnardo's (2023) Your Voice Matters 2022

5. Moderation systems and processes

Regulated services must have appropriate systems and processes in place to prevent and protect children from harmful content. These systems must address harm across all aspects of a service. Moderation systems and processes are in addition to, not a substitute for, ensuring services are safe and age-appropriate by design.

In its child safety duties codes and guidance, Ofcom must require that regulated service providers' moderation systems and processes have the following elements:

- **Internal moderation policies:** Written policies setting out the standards in sufficient detail that everyone who is involved in meeting those standards is clear about what is and isn't acceptable.

- **External policies:** It must be simple for children to understand what is and is not allowed on a service. Services must ensure rules (e.g. Community Standards or Guidelines) are easy to access and understand and presented in a way that is child friendly. Regulated services must set out what children can expect from them and how to complain if they fall short.

- **Automated moderation systems:** Where automation is used as part of a moderation system these models must make consistent and accurate determinations and be subject to human oversight. Moderation systems and processes must include strategies for early detection and responding to evolving harms, unexpected events and operate with capacity to moderate at speed and scale.

- **Moderators:** Those working in moderation teams must have sufficient training to be able to make consistent and accurate decisions. Service providers must conduct frequent sampling of decision-making to ensure that policies are being applied correctly and to identify moderators who are struggling so they can be offered further training or performance management. Moderators must also be offered pastoral care if they consistently come across material that may impact on them, and care must be taken to ensure they do not become desensitised.

- **Resourcing:** Safety functions in regulated services must be adequately resourced considering the level of risk and number of children who access it. If a service provider determines that it is disproportionate to have a dedicated safety function within its organisation, it must explain the basis for this decision in its risk assessment.

- **Trust and safety's mandate:** If a service has a dedicated trust and safety function, this function must have autonomy and authority to require measures necessary to ensure the safety of children. This includes autonomy to make changes and overrule or modify proposals from other teams whose goals may be in tension with safety (for example, product, marketing, public affairs, communications, and monetisation teams). Given the importance of trust and safety's work, trust and safety leadership should report directly to the CEO.

- **Product development review:** Internal safety experts must be included in the design and development of new products at an early stage. Their advice must be a determinative factor in deciding whether to launch a new product or any updates including features or functionalities. Safety experts must also advise on launch timings, testing phases, implications for moderation resourcing and any other capacity or feasibility issues.

- **External validation:** Decisions on safety must be made in consultation with fully independent, external experts and children themselves.

6. User reporting

Children often struggle with widely different and complex reporting processes on different services. Bereaved families who have lost children to online harms have also given testimony to how impossible it can be to find humans in the system to speak to when the worst has happened.

In its child safety duties codes and guidance, Ofcom must require that:

- All services, however small, must have a contact for complaints. Reporting and complaints systems must be consistently applied, easy to use, child-friendly, provide time-bound feedback, have safeguarding and support in place. Services must take steps to proactively bring user reporting mechanisms and systems to the attention of children.

- Reporting systems must enable children and trusted adults to complain about any aspect of a regulated service that is in scope of the children's safety duties in the Act. It must be possible for unregistered users to make a report and if the report is in relation to children's safety, a human must be involved early-on in the process.

- Once a service is aware that a child is involved, priority must be given to these reports, ensuring children are given heightened protections. Reporting and complaints systems must have a human early in the process where a child is involved.

- Regulated services must consider whether *Trusted Flagger* programmes can support children and their trusted adults (e.g. parents, carers, guardians, social workers, teachers, siblings) to raise issues about the service.

- If a regulated service provider takes action that impacts a child, it must communicate this in a way that is age appropriate, provide sufficient information so the child understands what has happened and why, and explains their right to appeal and the process for doing so.

- Regulated services must record data on user reporting and incorporate it into ongoing risk assessment. This information should be shared at the highest level.

User controls have a particular role to play in supporting children with certain characteristics or who are members of certain groups. However, they are not a substitute for providing a safe and age-appropriate service by design and by default.

In its child safety duties codes and guidance, Ofcom must require that:

- Regulated service providers must ensure that the user controls they offer provide children with the ability to manage the content they see and the way they use features and functionalities whilst still continuing to access the service.

- User controls must be available to all children and not be dependent on a parent or guardian activating parental controls.

If a regulated service provider includes media literacy as a mitigating factor in reducing the risk of harm to children, it must ensure that it is demonstrably effective and that it can evidence the same. General online safety initiatives delivered off platform are highly unlikely to be considered effective in mitigating platform specific risks to children and should not be included in children's risk assessments. Media literacy strategies must reflect the principles in Ofcom's *Media Literacy by Design: Best practice principles for on-platform interventions to promote media literacy*.[18]

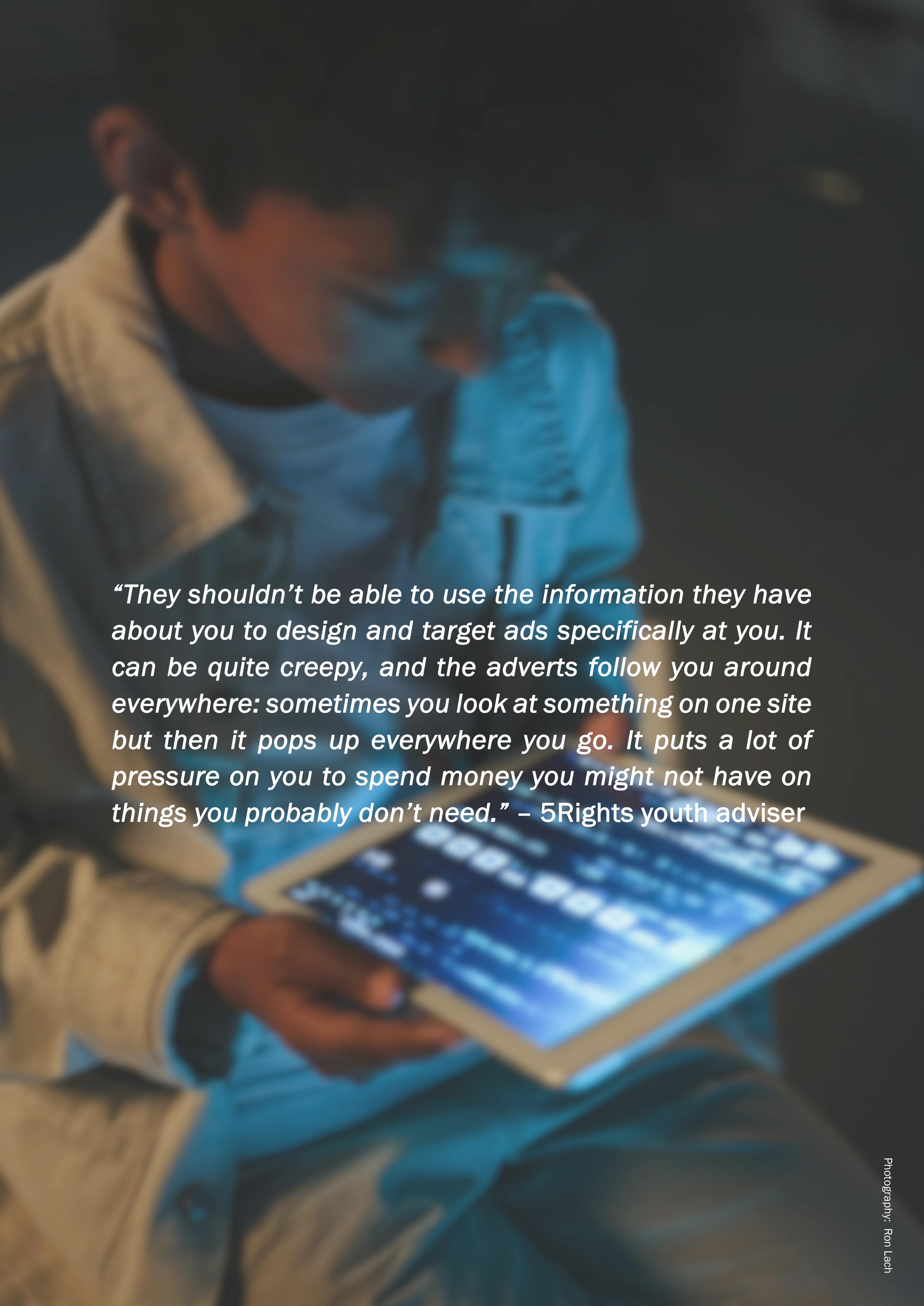
[18] Ofcom (2023) Media literacy by design: Best practice principles for on-platform interventions to promote media literacy

7. Terms of Service and transparency

Terms of service must include information about the steps providers are taking to protect child users from harm, including information on the presence of high-risk features or functionalities.

All terms, policies and community standards must be easy to find for children, concise, and in both format and language suitable to the age and diverse needs of children.[19] Non-textual messages, such as cartoons, videos, images, icons, or gamifications, can be helpful. Terms should also be presented in short and timely notifications along the user journey to ensure meaningful engagement, including at the point that specific options are activated.

[19] UNCRC General comment No. 25, Para. 39



"They shouldn't be able to use the information they have about you to design and target ads specifically at you. It can be quite creepy, and the adverts follow you around everywhere: sometimes you look at something on one site but then it pops up everywhere you go. It puts a lot of pressure on you to spend money you might not have on things you probably don't need." – 5Rights youth adviser

Photography: Ron Lach

8. Staff policies and practices

Trust and Safety (T&S) as a profession has not yet established minimum training requirements or professional standards, and those working in this field are not subject to oversight. Beyond whistle-blower protections, they do not enjoy heightened protections that allow them to operate within companies with the same autonomy and authority that other compliance functions enjoy.

In the absence of these key aspects of a professional framework, providers or regulated services should be encouraged to work within their organisation and across industry to:

- Codify the rights and responsibilities of T&S staff including a clear mandate to act independently in the best interests of users – even when doing so conflicts with the business interests of the service provider.

- Define a set of ethical principles to guide all aspects of T&S staff's work and the way they advise colleagues and leadership.

- Describe the minimum skills and expertise requirements for different functions within T&S and ensure that current and prospective workers meet these requirements or are given training to do so.

- Ensure skills, expertise, training, and support is consistent across all T&S staff and in equivalent roles, irrespective of whether they are employed directly or via a third-party service provider.

- Consider whether T&S team members should be able to participate in share option schemes and other incentivisation schemes that may create a conflict of interest and, if so, identify alternative schemes that incentivise enhancing safety standards and which are of equivalent value.
- Create a dedicated, confidential reporting channel so that any member of the T&S team who believes their ability to carry out their safety duties is being compromised can report their concerns without fear of retribution or penalty.

9. Governance and the business model

Governance is the codification of roles and responsibilities within organisations, it identifies decision-makers and accountabilities, it describes the process that must be followed to make decisions, and the values and rules that guide decision-making. Typically, the Board of Directors is the body with responsibility and accountability for ensuring sound governance.

To deliver for children, this regulation must spark a step change in the priorities of the tech industry – from the "move fast and break things" culture of the past to one where children's safety is front and centre. Creating services which are safe by design must become a responsibility that is held at the highest levels of these organisations and this must become a key measurement of compliance. Senior leadership must be made accountable for the adverse impacts their services and business models have on children.

Testimony from Meta whistle-blowers Frances Haugen[20] and Arturo Bejar[21] illustrates how service providers prioritise business growth at the expense of children, which has facilitated and exacerbated harms.

In its child safety duties codes and guidance, Ofcom must require that:

- Non-trivial design decisions and changes must be considered, approved, and recorded by senior leadership who must be provided with a comprehensive written briefing on potential risks and harms and the efficacy of any mitigation measures implemented before they decide whether to approve it.

- Senior decision-makers receive bespoke training to ensure they have the necessary skills to make informed decisions about whether the service is safe.

- The advice, leadership's deliberations and final decisions must be recorded in writing and form part of the regulated service's risk assessment. This reporting must have input from independent online safety specialists.

- If leadership rejects a recommendation from internal or external child safety experts on a non-trivial aspect of the regulated service provider's child safety strategy, this must be recorded in the risk assessment.

[20] The New York Times (2021, updated 2023) Whistle-blower says Facebook 'chooses profits over safety'

[21] The Wall Street Journal (2023) His job was to make Instagram safe for teens. His 14-year-old showed him what the app was really like

- Services must be transparent about their business model and describe it in their risk assessment, including identify aspects that are likely to increase or decrease risk of harm to children.

- Where risk is identified, service providers must adjust the business model or take demonstrably effective steps to mitigate the risks.

- Where a service has identified that a business model reduces risk, they should describe why and how the business ethos manifests in tangible safety measures.