

CY26 H1 Brazil ECA Digital (Digital Statute for Children and Adolescents) Transparency Report

Microsoft

Table of Contents

Introduction 3

A. Measures to Identify Children's Accounts and Unlawful Acts 4

B. Channels for Receiving Complaints and Investigations Procedures 5

C. Number of Complaints Received 9

D. Content and Account Moderation Instances by Type 9

E. Technical Improvements for Data Protection and Privacy of Children and Adolescents 10

F. Parental Consent Verification 11

G. Impact Assessment Methodology and Risk Management for Child Safety 14

Microsoft

Introduction

At Microsoft, we are committed to protecting children and adolescents online in Brazil and around the world. We recognize the importance of creating safe digital environments where young users can learn, connect, and explore without being exposed to harmful content or conduct. This commitment is reflected in our comprehensive approach to online safety, which combines advanced technology, clear policies, and ongoing collaboration with stakeholders including parents, educators, regulators, and civil society organizations.

This transparency report has been prepared in compliance with the requirements set forth in Law No. 15,211 of 2025 (Article 31, subparagraphs I through VII) and Decree No. 12,880 of 2026 (Article 45, subparagraphs I and II). The report covers the period from January 1, 2026, to June 30, 2026, and provides information about Microsoft's policies, practices, and outcomes related to the protection of children and adolescents across our digital services; for example, Bing, Microsoft Store, Xbox, and Teams (collectively referred to in this report as "Microsoft Consumer Services").

For almost two decades, Microsoft has been a global leader in protecting children online through its efforts to combat online child sexual exploitation and abuse ("CSEA"). Microsoft's approach is grounded in the recognition that this is a whole society challenge demanding coordinated action across government, law enforcement, technology providers, civil society, and survivor support organizations. Our work reflects a simple but nonnegotiable principle: children must be able to participate safely in digital services without being exposed to grooming, sexual abuse imagery, exploitation, or other harms. While this stance applies across all Microsoft products, the nature of risk differs widely by service. Our approach is therefore intentionally risk-based, service specific, and aligned with broader ecosystem responsibilities.

Reporting is a critical component of the global child safety ecosystem. As a U.S. headquartered company, Microsoft reports apparent child sexual abuse material ("CSAM"), grooming, and other required harms to the National Center for Missing and Exploited Children ("NCMEC"), which serves as the global clearinghouse for Cybertips, including delivering targeted Cybertips to the Federal Police in Brazil. We voluntarily publish Digital Safety Content Reports and service-specific reports such as the Xbox Transparency Report, which detail our enforcement actions, reporting volumes, and trends in abuse across services.

Finally, Microsoft recognizes that combating CSEA is not a static effort and requires ongoing multistakeholder collaboration. We prioritize transparency and user empowerment, including parents, educators, and young people themselves, to ensure that safety solutions are understood, trusted, and accessible. Building a safer digital

ecosystem for children and adolescents requires constant iteration, and Microsoft remains deeply committed to advancing this work in partnership with ANPD and the Brazilian and global community.

A. Measures to Identify Children's Accounts and Unlawful Acts

Identification of Children's Accounts

Microsoft employs multiple mechanisms to identify and appropriately manage accounts belonging to children and adolescents. Microsoft currently deploys a layered, ecosystem-wide age assurance framework that combines parent-child account creation at device or service set up, parent attestation of age, and parental consent and control flows. Together, these layers provide a consistent, privacy-respecting foundation for enforcing age-based protections across services.

Microsoft operating systems and services facilitate creation of a Microsoft Service Account ("MSA"). When a user sets up a Windows-based PC or an Xbox console, for example, an account must be created in order for set-up to proceed. Through information contained in set-up instructions and the set-up flows, parents are encouraged to create their own account first, and then, by acting as the "IT administrator" of the family, create and attach a child account to the parent account as part of a family group. When a user indicates that they are under the age of 18, Microsoft requires a parental consent process through which the parent confirms the child's age and authorizes their account. Once a parent authorizes the account, that account is flagged as a child account across Microsoft's ecosystem. This declared-age signal triggers age-appropriate experiences across many services, including service level restrictions, advertising limitations, heightened privacy protections, and access to parental controls through Microsoft Family Safety.

Parental consent flows serve both as an assurance mechanism and a governance tool. When a child attempts to create or use a Microsoft Account, Microsoft requires a parent or guardian to authenticate and approve the account. Parents then have a range of parental controls they may set through the Microsoft Family Safety suite of tools. These settings apply across Windows, Microsoft 365 Consumer services, Xbox, and other experiences, providing families with unified control.

Detection of Unlawful Content and Conduct

Microsoft uses a combination of automated technology and human review to detect unlawful content and conduct on our platforms, particularly those that may harm children and adolescents. For almost two decades, Microsoft has been a global leader in protecting children online through its efforts to combat online CSEA. Microsoft's approach is grounded in the recognition that this is a whole of society challenge demanding coordinated action across government, law enforcement, technology providers, civil society, and survivor support organizations.

At the foundation of our work is a comprehensive policy framework that prohibits the use of any Microsoft service to exploit or harm children. We maintain a strict zero tolerance stance on CSAM, grooming, and emerging forms of CSEA, including AI-generated material. These policies are enforced globally and apply across Microsoft's Consumer Services. They are complemented by robust enforcement processes that rely on both automated detection technology paired with human decision-making.

Detection Technologies

- PhotoDNA and Hash-Matching Technology: Microsoft developed PhotoDNA in 2009, in partnership with Dartmouth College, a globally adopted hash-matching technology that enables responsible platforms to detect and remove known CSAM at scale. Recognizing the ecosystem challenge, we subsequently made PhotoDNA freely available for licensing by law enforcement and industry worldwide. It is now used by hundreds of entities to detect and address CSAM. We have also invested in ongoing updates, including to address where changes have been made to an image and to improve speed, and most recently have made licensing freely available through the Tech Coalition. This same technology has also been licensed to StopNCII.org.

B. Channels for Receiving Complaints and Investigations Procedures

Microsoft's strong commitments to children safety online, informed by the principles of the ECA Digital, includes structured reporting channels and expedited procedures to ensure their best interest and progressive autonomy. Microsoft structures its systems for detecting, moderating, and responding to harmful content and conduct based on a preventive, risk-based approach guided by global digital safety standards, with a special focus on protecting children and adolescents from serious harm, such as sexual exploitation, grooming, violence, and other forms of online abuse. This approach integrates automated technologies, specialized human review, clear escalation flows, and accessible reporting mechanisms.

When it comes to detection and moderation, Microsoft employs advanced automated systems to identify CSEA material, combining the use of industry-shared hash databases—widely recognized as an international standard—with proactive classification and machine learning technologies capable of flagging content and behaviors associated with priority risks. These systems enable the rapid identification of known illegal content and the detection of new patterns of abuse, strengthening preventive capabilities even before user reports occur.

All end users must agree to the Microsoft Services Agreement and its associated Code of Conduct, which sets baseline rules prohibiting harms including illegal content, harassment, hate, exploitation, fraud, or attempts to bypass platform protections. The Agreement applies to user-generated content, user-to-user interactions, and content

generated via AI experiences. Service-specific rules, such as the Xbox Community Standards, build on the Agreement and tailor behavioral expectations for particular communities.

Microsoft also maintains a comprehensive content and conduct policy framework that governs all Microsoft Consumer Services. This framework uses a global taxonomy that defines harmful content categories and describes how each category should be addressed. All enforcement actions taken under these policies are designed to be proportionate, nondiscriminatory, and transparent. Microsoft documents how violations are identified, how enforcement decisions are made, and how users are informed of the outcomes. This approach is consistent with Article 33, § 2 of the ECA Digital. Our policies are publicly available on our Digital Safety site and we maintain a version for young people.

Resources and Education for Children and Adolescents

Microsoft recognizes that protecting children and adolescents online requires more than technological safeguards and parental controls—it also requires empowering young people themselves with knowledge, resources, and tools to navigate the digital environment safely and responsibly. Microsoft maintains dedicated resources designed specifically for children and adolescents to help them understand online risks and responsibilities when using Microsoft services. Our Digital Safety site for young people provides age-appropriate explanations of Microsoft's Code of Conduct, including: what behaviors and content are not allowed on Microsoft services; how to recognize harmful conduct such as bullying, harassment, and exploitation; and how to report concerns when they encounter problematic content or behavior.

Reporting Channels

Microsoft's Consumer Services feature "report a concern" mechanisms that allow users to flag inappropriate content or behavior. These mechanisms exist in-product, as well as in a centralized location experience. Reports feed directly into Microsoft moderation queues. Traditionally these features have enabled users to report based on commonly understood categories of content. This type of solution will also enable prioritization of categories of content deemed risky to children under the ECA Digital.

For high-risk harm categories such as child safety or violent extremist content, Microsoft maintains specialized escalation channels with NGOs, hotlines, and other trusted partners.

Investigation Procedures

Microsoft's approach to moderation and enforcement reflects a comprehensive, risk-based framework designed to protect children and adolescents from online harm

while respecting human rights, privacy, and freedom of expression. Microsoft uses a clearly defined moderation lifecycle across Microsoft Consumer Services. The lifecycle begins with the development of detailed user rules and content policies. Each service then conducts risk assessments to identify the harms that are most likely to occur on that platform. These risk assessments are incorporated into the design and deployment of detection technologies, the development of service-specific policies and enforcement protocols, and the training of human moderators ensuring that safety measures are tailored to the specific risks identified for each platform.

Microsoft uses automated detection technologies and user reporting mechanisms to identify content or behavior that may violate its policies. User reports and content flagged via automated detection that result in enforcement action may be subject to human review by trained moderators. When violations are confirmed, Microsoft applies appropriate enforcement actions and, where feasible, notifies users and provides an opportunity for appeal. Microsoft continuously analyzes safety data and updates its systems to ensure ongoing improvement in moderation effectiveness.

Detection and Review

Microsoft employs a multi-layered approach to content moderation across Microsoft Consumer Services that combines automated detection technologies, human review processes, and specialized operational hubs to identify and remove harmful content swiftly and accurately. Depending on the service, the multilayered detection system can consist of automated detection tools, together with rule-based systems (hash matching), heuristics and pattern detection, and human review to support the identification and handling of potentially violating content.

PhotoDNA and MD5 hashing technologies are deployed to detect known child sexual abuse imagery and terrorist content and we are in the process of adopting this technology for non-consensual intimate imagery on images and video content shared through Microsoft Consumer Services and on content uploaded for visual image searches of the internet or to support image or video generation by AI. Additionally, URL reputation systems in search evaluate websites and promote, demote, or block content based on policy and user settings. AI guardrails are also deployed to block unsafe inputs and outputs. These systems operate in near-real time to prevent the spread of harmful content or to route it for review.

Human Review and Escalation

Microsoft employs specialized moderation teams trained to handle harassment and abuse, self-harm and crisis concerns, child sexual exploitation content, terrorist or extremist content, false reports and report-abuse patterns, context-dependent cases (e.g., satire vs. threats), and appeals of enforcement decisions. Sensitive cases, especially those involving children, real-world harm, or legal obligations—are escalated

to highly trained staff and reported to authorities when necessary. For example, when PhotoDNA generates a match for the first time, at least two moderators independently review the content to determine if it can be confirmed as CSAM before enforcement action is taken.

Reported or detected content is routed through a structured human review workflow. Reviewers evaluate the content against applicable platform policies which include local legal requirements. Based on this assessment, the content is classified and an enforcement outcome is determined. Classification outcomes include no policy violation, policy violation with graduated enforcement, or case escalation. Upon escalation, Microsoft may consider user history to determine actions to be taken.

For content that is confirmed to violate policy or law, including the ECA Digital, Microsoft applies proportionate enforcement actions, as further described below. Where legally required in cases involving child sexual exploitation, content and associated information are escalated to appropriately trained specialist teams and reported to NCMEC. All enforcement actions are logged and subject to ongoing quality review. Affected users may access applicable appeal mechanisms, and appealed decisions are reassessed by human reviewers not involved in the original determination.

Enforcement Actions

Microsoft applies a graduated set of enforcement actions intended to be objective, proportional to the nature and severity of the harm, and consistent with applicable legal obligations, including protection of the rights of children and adolescents. Enforcement measures may include content-level actions, such as removal, blocking, restriction of visibility, or disabling of sharing functionalities where content violates Microsoft policies or applicable law, such as the ECA Digital. Depending on context, enforcement may also include account-level or feature-level actions, including warnings, temporary restrictions on functionality, limits on communications, time-limited suspensions, or account closure in cases of repeated or egregious violations.

Enforcement decisions consider factors such as risk of harm, recurrence, user behavior, and legal requirements, and are subject to human review in high-risk or sensitive scenarios, including matters involving children and adolescents. Microsoft also employs safeguards to address abuse of reporting and complaint mechanisms, including measures to limit or suspend reporting capabilities where complaints are demonstrably malicious, automated, or manifestly unfounded, in order to preserve the integrity and availability of reporting systems for legitimate safety concerns.

Notice and Appeals

Where feasible and appropriate, Microsoft provides users notice of moderation enforcement actions and access to an appeal mechanism. Enforcement notices typically explain the action taken and reference the applicable policy basis for the action. Users may request a review of certain enforcement decisions through established appeal pathways. For example, users can access information regarding appeals here[1].

Response Timeframes

Microsoft seeks to identify and remove harmful content swiftly and accurately. Reported content is triaged through automated prioritization and human review queues, with imminent harm signals including credible threats, self-harm crisis indicators, child sexual exploitation and abuse, and terrorism prioritized for expedited review. Standard policy violations are reviewed within defined internal service levels calibrated by volume and risk, with enforcement actions applied promptly following classification. Appeals are reviewed within targeted timeframes by reviewers not involved in the initial decision. Response performance is continuously monitored through quality assurance and operational metrics, adjusted to account for spikes in reporting or public safety events, and documented to support transparency and regulatory oversight.

Government reports, lawful requests, and judicial or administrative removal orders, are processed through dedicated channels, assessed for legal validity and scope, and addressed within the timeframes specified under applicable law or order terms. Where orders require timed or immediate removal, Microsoft executes enforcement promptly and documents actions taken.

C. Number of Complaints Received

During the reporting period of January 1, 2026 to June 30, 2026, Microsoft received 101,379 complaints from users in Brazil across the Microsoft Consumer Services covered in this report.

D. Content and Account Moderation Instances by Type

During the reporting period, Microsoft took the following moderation actions across Microsoft Consumer Services in Brazil. These actions resulted from both proactive detection through automated systems and reactive responses to user reports.

Moderation Actions by Type:

Measures Taken on Content – Content Deleted				
Harm Type	Bing	Microsoft Store	Xbox Studio	Teams Consumer

[1] Digital Safety | Moderation and Enforcement | Appeals

Bullying and Harassment	-	6	108,187	100
Child Sexual Exploitation and Abuse	6,191	-	3	198
Grooming	-	-	504	1
Hateful Conduct	-	23	21,967	73
Inappropriate Disclosures of Sensitive Personal Information	-	-	28,206	21
Suicide and Self-Injury	-	-	3,766	-
Violent Extremism and Terrorism	3	-	293	-
Violent Threats, Incitement, and Glorification of Violence	-	4	24	10
Total	**6,194**	**33**	**162,950**	**403**

Measures Taken on Account – Account Suspension				
Harm Type	Bing	Microsoft Store	Xbox Studio	Teams Consumer
Bullying and Harassment	-	-	983	87
Child Sexual Exploitation and Abuse	-	-	3	137
Grooming	-	-	7	1
Hateful Conduct	-	-	807	67
Inappropriate Disclosures of Sensitive Personal Information	-	-	-	-
Suicide and Self-Injury	-	-	-	-
Violent Extremism and Terrorism	-	-	-	-
Violent Threats, Incitement, and Glorification of Violence	-	-	-	11
Total	-	-	**1,800**	**303**

E. Technical Improvements for Data Protection and Privacy of Children and Adolescents

During the reporting period, Microsoft continued to maintain and develop technical and organizational measures designed to protect the personal data and privacy of children and adolescents.

Microsoft uses data protection impact assessments ("DPIAs") as part of its privacy governance mechanism to identify, evaluate, and mitigate privacy and data protection risks associated with personal data processing. Where a DPIA is required and the relevant product, service, or feature is likely to be used or accessed by individuals under

18, Microsoft's DPIA framework requires consideration of child-specific factors including the child's best interest, privacy and safety, child-specific controls, and risks that may affect children differently. Microsoft maintains a unified, enterprise-wide DPIA framework with assessments tailored to the particular service and processing activity.

Microsoft uses complementary privacy and digital safety assessment processes to identify and mitigate risks affecting children and adolescents. Privacy impact assessments focus on risks arising from the processing of personal data and, where relevant, include consideration of the best interests, privacy and safety of children, child-specific controls, and risks that may affect children differently. These assessments are complemented by Microsoft's Digital Safety risk-assessment processes, which address broader risks to the safety and well-being of children and adolescents arising from the design, functionality and operation of digital services. Together, these processes support Microsoft's identification and management of privacy, safety and health risks relevant to children and adolescents, including for purposes of applicable requirements under the ECA Digital.

Windows disables and locks several device-based data-collection features for Microsoft child accounts for users under 18, including the Advertising ID, optional diagnostic data, and personalization/profiling features. Microsoft does not deliver personalized advertising to children whose birthdate in their Microsoft account identifies them as under 18 years of age. Location services and Find My Device for child accounts are also disabled and locked in order to limit collection of child location data. Child accounts are also blocked from participating in Feedback Hub and product surveys that collect personal information.

F. Parental Consent Verification

Through Microsoft Family Safety[2], the Microsoft Service Account ("MSA") identity implements parental supervision and child-safety protections anchored to linked parent/guardian accounts and child accounts. These controls apply consistently across devices leveraging the Windows operating system, Xbox consoles, and services using the MSA identity solution.

Protections follow the child's account. Once the child signs into an MSA, rules can be enforced across the MSA ecosystem, including Windows-enabled personal computers. This approach mirrors the administrator–end user access patterns used in workplaces and educational institutions and provides the same predictability and auditability that supports compliance outcomes as one might see in an organizational context. By connecting age signals, parental authorization, and safety controls through a persistent account-based identity, the model provides a practical mechanism for

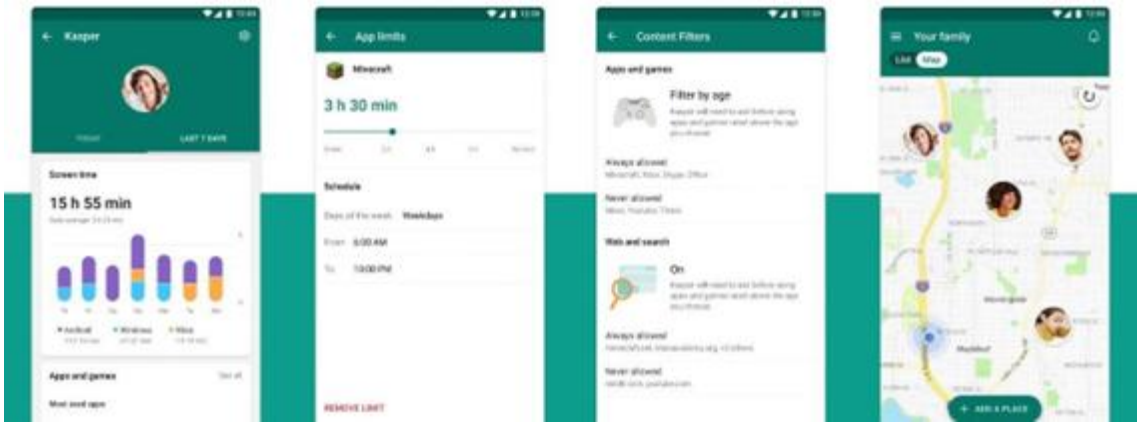
[2] Microsoft Family Safety | Microsoft 365

translating the legal requirement for heightened protection of children and adolescents into consistent, enforceable product-level safeguards

Identity-Based Supervision Tools

Microsoft offers a comprehensive, integrated, and interoperable set of parental supervision tools distributed across its main digital ecosystems, including Microsoft Family Safety, Xbox Family Settings, and family groups linked to MSAs. These tools are designed to enable parents and guardians to exercise effective, continuous, and informed control over the use of online services by children and adolescents, respecting the principle of comprehensive protection and recognizing the central role of the family in mediating the digital experience of minors.

The architecture of these solutions is based on individual accounts and linked profiles, which allows for the application of different settings according to the user's declared age, level of maturity, and the choices of their guardians. Through Microsoft Account family groups, parents and guardians can centralize the management of children's and adolescents' digital experiences, setting usage parameters that apply consistently across different devices and services, including operating systems, games, applications, and communication platforms.



Microsoft Family Safety App

Parental supervision tools offer granular and configurable features, including, among others, account privacy and data management, purchase and financial transaction restrictions, control of contacts and social interactions, setting screen time limits, age-appropriate content filters, and access to activity reports. These features allow guardians to transparently monitor usage patterns, identify potential risks, and adjust settings as the child or adolescent's needs and development evolve.

This model reinforces a protection approach based on active supervision and informed participation by guardians, rather than relying solely on intrusive age verification mechanisms or generalized assumptions about users' ages. By empowering legal guardians with effective and accessible tools, Microsoft seeks to mitigate specific

risks of exposure to inappropriate content, harmful interactions, or abusive behavior, while preserving user privacy, and service usability.

These solutions are developed according to principles of safety, security and privacy by design, with clear interfaces, accessible information, and no manipulative practices that could undermine parental supervision. In addition, the tools are continuously improved in light of risk assessments, user feedback, technological advances, and regulatory guidance, ensuring that they remain effective, proportionate, and aligned with national and international best practices for protecting children and adolescents in the digital environment.

a. Family groups and parent–child roles

Parents create a family group and add child accounts to that group. From this structure, parents can configure and administer parental controls. Parent accounts possess administrative privileges that enable them to manage limits and permissions, consent, web/search filters, control app/game access, purchases, screen time, and activity reporting.. The architecture of this identity solution supports parental consent and parental date of birth validation. Windows' setup guides parents to create or link a child account and activate parental controls.

b. Cross-device enforcement of rules

When a child logs in, the controls set by the parent apply on devices leveraging the Microsoft Account, including personal computers running Windows. Protections remain intact regardless of whether the child uses a newly purchased PC, a shared household laptop, or a school use device configured for personal use at home, as long as the same identity mechanisms are leveraged. Importantly, the cross-device identity does not allow cross-service data sharing for purposes of, for example, profiling, advertising, or creating behavioral analytics of minors. Microsoft enforces strict internal boundaries to prevent such uses.

c. Transparency and auditability

Parents receive optional activity reports and at any time can review the controls applied to the child's account. This transparency gives families meaningful oversight and enables the company to audit consistent application of the protections under a unified identity structure.

d. Content Filtering and Access Restrictions

Microsoft Family Safety offers strict content filtering defaults to Microsoft child accounts. These settings also facilitate content filtering by default on the Microsoft Store based on the child's age. Through Microsoft Family Safety, parents can add additional restrictions by choosing apps, including games, to filter based on age-appropriateness, ensuring kids only access content suitable for their age. Parents set

the maturity rating limit for each child (for instance, only allow apps and games rated for ages 10 and under). Apps and games rated above the set age limit will require parental approval before download. This mechanism keeps kids safe by preventing exposure to violent, sexual, or other inappropriate content based on age rating, while still giving families flexibility.

e. Purchase Restrictions

Microsoft Family Safety offers parental controls to prevent unauthorized or excessive spending by kids, such as on games or in-game purchases. Parents can require that any purchases the child attempts must be approved by an adult first. Even for free downloads, this gives parents a chance to vet content before the child accesses it. Instead of linking a credit card for open spending, parents can also set up an allowance by adding a fixed amount of money to the child's account. The child can spend this pre-loaded balance but cannot exceed it, effectively capping how much they can purchase on their own. This prevents surprise credit card bills and teaches kids about spending limits.

f. Screen Time and Activity Monitoring

Through Microsoft Family Safety, parents can set screentime schedules, daily time limits for apps and games, as well as daily time limits on devices such as Windows PCs or Xbox consoles and receive activity reporting about their child's usage across devices to which the child is logged in using the MSA. Parents can define exactly how much time per day a child is allowed to use specific apps or devices and even set scheduled "curfew" hours. Once set, the child account is automatically signed out or blocked when their allotted time is up. Parents can adjust these limits, and children can also request additional time which the parent may approve remotely on a one-time basis. Microsoft Family Safety also provides weekly activity reports for each child in the family group, which give parents a clear picture of the child's online activity. These reports summarize things like which apps or devices were accessed and for how long, and attempted purchases. The report also highlights the top apps used by the child during the week. Microsoft also provides guidance to parents on how to use these features as part of a holistic approach to digital parenting, such as using activity reports as conversations starters with a child.

G. Impact Assessment Methodology and Risk Management for Child Safety

Impact Assessment Methodology

Microsoft welcomes the adoption of risk assessment as a central regulatory tool, as it enables a more proportionate, evidence-based, and preventive approach to online child safety. In this context, Microsoft has a structured methodology through

which its services can systematically identify, evaluate, and mitigate risks to children and adolescents arising from the design, functionality, and operation of their services.

Microsoft adopts a structured framework of risk assessments on a global scale, designed to identify, analyze, and mitigate risks associated with online harm, both for the general public and, specifically and reinforced, for children and adolescents. This framework is based on the systematic identification of relevant categories of harm, such as child sexual exploitation and abuse, grooming, harassment, violence, exposure to pornography content, risks associated with social interactions, and misuse of personal data, as well as risk vectors related to features, interaction flows, service architecture, and user behavior.

Based on this analysis, proportionate mitigations are identified, which may include technical safeguards, design adjustments, access controls, moderation mechanisms, age-based configuration standards, parental supervision tools, and educational interventions. This process is applied consistently across different products and services, respecting their technical and usage specificities, and is integrated into the development lifecycle, from the design phase to the ongoing operation of services. This approach reflects the reality that digital services are not "one-size-fits-all." Instead, Microsoft tailors safety interventions to the risks and responsibilities of each service and feature.

Safety by Design Reviews

Microsoft Services are required to complete a safety-by-design review prior to any substantial change or update to a service, reinforcing early identification of design risks before launch or major iteration. A safety-by-design review examines how a new or evolving service enables user interaction or content creation, anticipates how the design of a service might be used toward harmful ends, and evaluates whether appropriate safeguards are built in from the outset. This includes assessing the nature of the feature, the types of users who may engage with it, and the potential for misuse or unintended consequences.

In practice, a safety-by-design review includes:

- Assessment of how the feature works and how users will engage with it, including whether it introduces new interaction patterns, incentives, or visibility surfaces that could generate digital safety risk.
- Evaluation of potential harms and risk vectors, using established frameworks to identify illegal or harmful content and conduct risks to children, risks amplified by automated systems, and other relevant harm categories.

- Review of child access and child safety exposures, which is required unless the service provides evidence to rebut the presumption of potential child access as outlined in Microsoft's risk assessment guidance.
- Documentation of proposed mitigations, including content policies, enforcement mechanisms, reporting pathways, automated detection systems, rate limiting controls, or product design adjustments.

The safety-by-design review is therefore part of a broader lifecycle approach to digital risks in the product. By completing these reviews before significant updates, Microsoft ensures that safety considerations are embedded deeply and consistently into the shaping of new features, and that services remain aligned with internal standards and evolving global regulatory expectations.