Exploring the Safety by Design Assessment Tool Modules



RMIT UNIVERSITY

Australian Government

eSafety Commissioner

Exploring the Modules

Let's take a look at what each of the assessment tool modules addresses, considering the purpose of each and the information they provide. Factors include improving business operations, and designing, developing and releasing online products that have user safety at their centre. The titles for the modules are listed in the diagram below.





Online Harms Module

The online harms module considers the complex factors that may be involved when using the online environment. Having in-depth knowledge of the different types of online harms and their impact can help the designer better build user safety and rights into their product or service.

The module provides information for the participant in relation to:

Background – Online Harms

There is general consensus that online risks and harms relate to both content (i.e. production, distribution and availability of content) and activity (i.e. consumption of content). Online risks and harms can also occur as a result of access, exposure and interaction with others.

The Impacts of Online Harms on Victims

Since Safety by Design focuses on the safety and rights of users, this module frames online harms through a human rights lens wherever possible, and emphasises the impacts of these harms on users.

Online harms can be extremely serious, with severe social, emotional, psychological and even physical impacts. This section of the module looks at the personal safety impacts, health and wellbeing impacts, emotional and social impacts, and financial impacts.

Typology of Online Harms

Online risks and harms are multi-layered and multi-dimensional. Each type of online harm is not exclusive; an incident may involve multiple types of harms and there may be overlap of typologies. This section of the module unpacks the different types of online harms:

- Personal safety harms that predominantly violate the right to personal safety
- Health and wellbeing content or activities that predominantly violate a person's right to health or wellbeing
- Deception and manipulation
- Harms that violate a person's right to participation or free expression
- Harms that violate dignity
- Harms that involve discrimination
- Online activities or content that predominantly interfere with a person's right to privacy.

Techniques and Tactics – Modes of Online Abuse

This section provides useful information about the different modes of online abuse and the techniques and tactics that are commonly used. It examines:

- Account types and uses
- Data, privacy and security
- Threats, manipulation and deception

Contributing Factors – Who May Be at Risk

This module section examines the factors that increase a person's risk online, including social and cultural factors. It considers intersectionality and online harms, as well as the disproportionate way that some groups are abused online.

The section considers victim risk factors and protective factors.

As well as victim factors, the section examines perpetrator factors.

NOTE: There are no questions for the participant in this module, nor is a tailored report generated for this module.

RMIT UNIVERSITY
Australian Government
eSafety Commissioner

Structure and Leadership Module

This module examines how embracing online safety as part of the organisation's culture can help to proactively address online harms.

The module provides practical information and further questions for the participant in relation to:

Corporate Values and Company Principles

The first section of this module highlights the importance of including reference to user safety in corporate values and company principles. The section includes information and examples on corporate values and company principles that reference user safety and/or Safety by Design.

Accountable Individual or Team

This section looks at team/individual responsibility for user safety. It includes information on a Safety by Design business model canvas as a means of visualising the value of integrating user safety and rights into the organisation's value proposition.

Board or Leadership Team

This module includes questions on the board or leadership team, as well as using key user safety data and measuring it against internal safety processes.

Employee Training

The employee training section examines:

- The importance of having user safety training in place for staff.
- Safety awareness program for employees on user safety and safety by design.
- Providing training and awareness for employees on how to report illegal content.

The section includes useful information on:

- Employee user safety training at induction, as well as ongoing training, and monitoring.
- Suitable training topics on user safety.

Investment and Reporting

This section includes questions on:

- Outlining user safety investments, innovations and operating costs clearly within the company's financial reports.
- Investing in internal or operational user safety processes.
- Reporting on user safety investments, innovations and operating costs.

Global Alliances and Memberships

This section examines the value of membership of global online safety alliances or initiatives. It provides information on fostering collaboration through global leadership.

RMIT UNIVERSITY
Australian Government
eSafety Commissioner

Internal Policies and Procedures Module

This module addresses how online safety needs to be embedded in internal policies, processes and procedures to help keep users safe. It considers how online harms can be clearly understood, assessed and addressed at each stage of the product design and development process.

The module provides practical information and further questions for the participant in relation to:

Internal Safety by Design Policies

This section asks questions about the documented Safety by Design processes that the organisation has in place, employee training on Safety by Design processes, and the individuals or teams responsible for developing Safety by Design processes.

Acceptable Use Policy

This section focuses on the acceptable use policy and human rights and children's rights principles, guidance, standards or frameworks to be considered when drafting the policy.

Employee Safety Practices

This section asks questions about the organisation's employee safety practices in relation to communication and use of the acceptable use policy, as well as of evaluation of the policy.

Formal Safety Review Process

This section asks you to consider whether formalised safety processes are part of the product or service design process.

User Registration or Sign Up

User registration or sign up is a key part of user safety – this section focuses on what information is captured at this point in the user experience.

Risk Thresholds and Profiles

Having clear risk thresholds and comprehensive risk profiles in place for content, conduct and environment on the product or service helps keep users safe. This section asks several questions about the establishment, type and use of risk thresholds and risk profiles by the organisation. The section includes a useful diagrammatic explanation of the factors that should be considered in conducting risk thresholds and risk profiles.

Internal Operation Guidelines

This section provides useful guidance on establishing internal operational guidelines which illustrate how the product or service is to be used in practice.

Safety Risk and Impact Assessment

This section deals with how safety risk and impact assessments are documented and used in the organisation and what factors the assessment addresses. It outlines the eight steps to carry out a safety risk and impact assessment. The section also considers whether safety, security and privacy by design features are given equal priority.

Implementing Safety by Design

This section asks questions about implementing Safety by Design in the organisation, particularly in comparison to security and privacy by design processes.

Internal Processes

This section deals with having internal mechanisms or processes available for employees to raise concerns about human rights, ethical or safety issues related to an online product or service.

Duty of Care – Employee Wellbeing

This section is about the support services that the organisation has in place to support employees and contractors who view or interact with illegal or harmful content.



Moderation, Escalation and Enforcement Module

A proactive approach to detecting, moderating and reporting online harmful content and activity is required to help ensure user safety. This module explores the practices, pathways and processes that an organisation should have in place to act on violations effectively and efficiently.

The module provides practical information and further questions for the participant in relation to:

Harm Prevention and Detection Processes

The first section of the module focuses on harm prevention, detection and reporting processes. It looks at:

- Tools for detecting, moderating and reporting illegal or harmful content or behaviour, including content that is not age-appropriate.
- The types of action the organisation takes to minimise exposure to harmful content or behaviour.
- What features of the product or service detection, moderation and reporting tools are applied to.
- What types on online harms the detection, moderation and reporting tools identify.

Service Monitoring Practices

This section looks at service monitoring practices such as:

- How often monitoring occurs.
- Whether the outcomes of monitoring practices are fed into system improvements.

Moderation Practices

This section looks at moderation practices such as:

- Whether moderation practices consider the characteristics of users.
- Providing prompts, incentives and rewards to encourage positive user behaviour.

Standard Operating Procedures for Illegal Content and Crisis Management

Questions and information in this section focus on:

- Whether the organisation has crisis management or online incident procedures in place.
- Standard operating procedures for reporting illegal content to law enforcement authorities.

Triage and Escalation Process

Questions and information in this section focus on:

- Training teams and/or individuals dealing with safety reports, complaints and violations.
- Who is involved in internal escalations.
- Triage and escalation processes to deal with violations of user acceptance policy or community standards.
- What levels of priority are applied to different categories of abuse when triaging or escalating cases.

Incident and Response Action

Questions in this section relate to:

- Actions undertaken as part of formal incident response plan for security, privacy and online safety incidents.
- What information the incident response register document contains.
- Timeframes for responding to incidents.

Enforcement

Questions in this section relate to:

- Methods of enforcing breaches to community standards, terms or service, or acceptable use policy.
- Ensuring users who have been removed or blocked are not able to re-register.



User Empowerment

This module focuses on how the rights and safety of users can be embedded into online services and products and how tech companies can amplify and strengthen user safety.

The module provides practical information and further questions for the participant in relation to:

Community Standards

The community standards section asks questions about the organisation's approach to translating acceptable use policy into plain language and short form notices for users.

The section includes practical considerations for developing community guidelines/standards that define a set of norms and expectations and code of conduct for users of the product or service.

Default Settings

The questions in this section focus on:

- Providing default safety settings to the highest levels at registration or sign up.
- Setting defaults according to the age of users.

Social Contracts

This section asks questions about the implementation of social contracts at the point of registration – i.e. the duties and responsibilities of the service provider, user and any third parties to help ensure the safety of all users. Examples of social contracts are also provided.

User Options/ Technical Tools

This section focuses on:

- Providing technical tools to allow users to manage their online experiences.
- Providing reporting features for users.
- User ability to choose the safety settings that are applied.

Safety Features

This section contains questions on communicating and explaining to users the use of automated systems, artificial intelligence, natural language processing or machine learning.

Reporting Mechanisms

Reporting mechanisms are a crucial part of ensuring user safety. This section asks questions about:

- The reporting mechanisms available for users to lodge complaints and how these mechanisms cater to the different capabilities and capacities of users.
- How reports are reviewed.
- Dedicated reporting channels available to users.
- Whether reporting mechanisms allow users to provide context to their reports.
- Putting in place steps to ensure users are not confronted by material that they have reported.
- Separate reporting systems for governments, trusted notifiers, or experts.
- Timeframes to action and resolve complaints.

Appeals Process

This section focuses on having a process in place that gives the user the ability to contest or appeal a complaint, removal or block.

Feedback Loop

This section asks questions about the type and frequency of feedback to users about the outcome or status of their report or complaint.

User Safety

The final section asks about whether/how the organisation directs users to external support or advice services for safety risks and harms.



Transparency and Accountability Module

This module examines the steps tech companies can take to ensure transparency and accountability are central to their operations.

The module provides practical information and further questions for the participant in relation to:

Public Information

The public information section focuses on organisations:

- Making information on safety policies and standards publicly available.
- Making information and statistics on safety processes publicly available.
- Having a dedicated section that houses public information.
- Ensuring safety policies and standards are accessible, inclusive, easy to understand, and available in multiple languages.

This section contains information about Trust and Safety Centres, addressing the needs of different users in publishing information, linking to external support services.

Informing Users

This section is about providing advice to the public about safety measures and policies and informing users about changes to safety measures and policies.

Measurement

This section deals with what is involved in measuring the accuracy and effectiveness of safety features, and the impact of safety features on users.

Internal Review

This section asks questions about the regular review of safety policies and standards, and communicating and implementing review findings.

Independent Review and Evaluation

This section asks questions about the organisation's:

- Engagement with external stakeholders in the development, interpretation, application and evaluation of safety policies.
- Consultation with different age groups in the development, interpretation, application and evaluation of safety policies and procedures.
- Consultation with diverse groups regarding safety policies.
- Frequency of user safety evaluations.

External Mechanisms

The external mechanisms section poses questions about the processes for people outside the organisation to raise concerns about human rights and ethical or safety issues.

Publication

The final section of this module asks about publicly publishing reports or data relating to user safety.

RMIT UNIVERSITY
Australian Government
eSafety Commissioner