SAFETY BY DESIGN

Safe Approaches For Ethical Technology *By Design*:

Making Products and Services Safer for Users and Consumers

CURRICULUM AND FACILITATOR GUIDE

October 2025

SAFETY by Design reflects the collaboration and contribution of many people and organizations engaged in preventing, responding to, and mitigating online harms. All sources have been cited.

Prepared by IREX with the safety-by-design expertise of Eugene Odanga Masinde, reviews and feedback from Development Gateway, and professional graphic design of Tamar Gabisonia. The team would also like to thank the experts at Australia's eSafety Commission for their invaluable guidance and feedback.

Date of publication: October 2025

TABLE OF CONTENTS

LIST OF ACRONYMS

AI	Artificial Intelligence
AR	Augmented Reality
CCDH	Center for Countering Digital Hate
CMCA	Computer Misuse and Cybercrimes Act (Kenya)
CRC	Convention on the Rights of the Child (UN)
CRbD	Child Rights by Design
CSEM	Child Sexual Exploitation Material
CSO	Civil Society Organization
DPA	Data Protection Act (Kenya, 2019)
E2EE	End-to-End Encryption
GDPR	General Data Protection Regulation (EU)
ICT	Information and Communication Technology
IoT/ IIoT	Internet of Things/ Industrial Internet of Things
KE-CIRT/CC	Kenya Computer Incident Response Team Coordination Centre
KPI	Key Performance Indicator
ML	Machine Learning
NC4	National Computer Cybercrimes Coordination Committee (Kenya)
NCII	Non-Consensual Intimate Imagery
NCMEC	National Center for Missing and Exploited Children
NCSC	National Cyber Security Centre (UK)
ODPC	Office of the Data Protection Commissioner (Kenya)
OECD	Organisation for Economic Co-operation and Development
OFCOM	Office of Communications (UK)
OSA	Online Safety Act (UK)
OWASP	Open Web Application Security Project
PDLC	Product Development Lifecycle
PIA	Privacy Impact Assessment
PII	Personally Identifiable Information
SbD	Safety-by-Design
TFGBV	Technology-Facilitated Gender-Based Violence
TSPA	Trust and Safety Professionals Association

TRAINING OVERVIEW

SCOPE OF TRAINING

This training program is designed to equip mid-level technology-based product professionals with knowledge and skills to effectively and sustainably implement and integrate safety-by-design (SbD) principles throughout their product life cycles. It focuses on their ability to proactively embed user safety to prevent and mitigate various forms of digital harms, especially harms targeting women and girls; enabling a safety-first culture; and preparing participants for hands-on prototyping, SbD features testing, and deployment.

OVERALL TRAINING GOALS AND OBJECTIVES

- **Foundational Understanding:** Define safety-by-design (SbD) and articulate its importance for product teams, end users, and broader society and consumer base within a given context.
- **Contextual Awareness:** Identify common digital harms, with a detailed focus on Technology-Facilitated Gender-Based Violence (TFGBV) and harms to women and girls, including types, prevalence, and significant impact.
- **Lifecycle Integration:** Map SbD activities across the Product Development Lifecycle (PDLC) and understand processes for proactive Risk Identification, Mitigation, and Prevention, applying these specifically to TFGBV risks.
- **Practical Design Skills:** Apply SbD principles to tangible design choices, analyze examples of safer features, understand the role and risks of safety technologies like Artificial Intelligence (AI)/Machine Learning (ML), and apply trauma-informed and survivor-centered design principles.
- **Implementation and Culture:** Understand strategies to cultivate a robust safety-first organizational culture, identify practical steps for SbD adoption while overcoming barriers, and grasp the principles of ethical co-design with users.
- **Anticipate Future Harms:** Take a broad look into the current tech ecosystem as well as the new generation of technologies such as AI, Virtual Reality (VR) and Augmented Reality (AR), and Internet of Things (IOT) in an attempt to anticipate harms that may arise as a result of the mass adoption of such tools and services.
- **Application:** Prepare participants to apply the learned concepts directly to their own digital products, services, and challenges in subsequent hands-on prototyping workshops.

TRAINING STRUCTURE:

Modules

Module 1: Introduction to Safety-by-Design

Module 2: Core Principles and Frameworks

Module 3: From Principles to Action: Operationalizing SbD

Module 4: Deeper Dive into Designing for Safety: Principles and Practices

Module 5: Building SbD Culture

Module 6: Emerging Technologies and Workshop Bridge

TARGET AUDIENCE

25-30 representatives (Product Managers, Designers, Engineers, Policy & Safety Staff) from commercial tech platforms, tech startups, and stakeholders.

FORMAT

A 2-day in-person training workshop, structured into six (6) interactive modules utilizing adult learning principles (discussion, case studies, group activities, simulations).

DURATION

12 hours – 6 hours per day

GENERAL FACILITATOR GUIDANCE

Preparation

Thoroughly review this facilitator guide, the accompanying slide deck, and all participant materials well in advance. Familiarize yourself with the key concepts, activities, and estimated timings. Where possible, prepare or research relevant local examples to supplement those provided.

Engagement Focus

While this guide provides detailed content, aim for an interactive and participatory session. Use this guide as a foundation and reference, but use the slide deck, encourage discussion, draw on participant experiences, and facilitate activities dynamically rather than relying solely on lecturing. Adapt your delivery based on the group's engagement and prior knowledge.

Time Management

Adhere to the estimated durations provided for each module and activity as closely as possible to cover all material. **Consider appointing a volunteer participant at the start of Day 1 to gently signal time cues** (e.g., 5 minutes remaining for an activity) to help keep the training on schedule.

Capturing Insights

Make active use of flip charts or whiteboards to capture key definitions, participant contributions during discussions, brainstorming outputs, and activity takeaways. This visual reinforcement aids learning and retention.

Day 1 Wrap-up & Day 2 Recap

At the end of Day 1, briefly summarize the key topics covered. **Consider asking for a participant volunteer to jot down 3-5 main takeaways from Day 1 and prepare to share them in a brief (2-3 minute) recap at the very beginning of Day 2.** This aids reinforcement and provides a peer-led start to the second day (complementing the planned recap activity).

GENERAL FACILITATOR GUIDANCE

Managing Questions & Discussion

Explicitly create space for participant questions throughout the modules, not just during activities. Gently guide discussions to stay relevant and manage time effectively, ensuring diverse voices have an opportunity to contribute.

Contextual Relevance

Consistently link the concepts, harms (especially TFGBV), and potential solutions back to the specific local context whenever possible, using local examples or prompting participants to share their relevant observations.

Flexibility

This guide is comprehensive, but feel free to adjust the depth on certain topics based on the participants' specific roles, industries represented, and expressed interests, while ensuring all core learning objectives are met.

MODULE

1

INTRODUCTION TO SAFETY BY DESIGN

MODULE 1

Introduction to Safety-by-Design

Learning Objectives

Upon completion of this module, participants should be able to:

1. Define safety-by-design (SbD) and its core concepts (proactive vs. reactive).
2. Articulate the importance and benefits of SbD for users, companies, organizations, investors, and society within their local context.
3. Identify the general landscape of technology-facilitated harms.
4. Define Technology-Facilitated Gender-Based Violence (TFGBV), its common forms, and its disproportionate impact.
5. Recognize the need for context-specific SbD approaches.

Materials Needed

- Participant handouts
- Markers
- Flipcharts

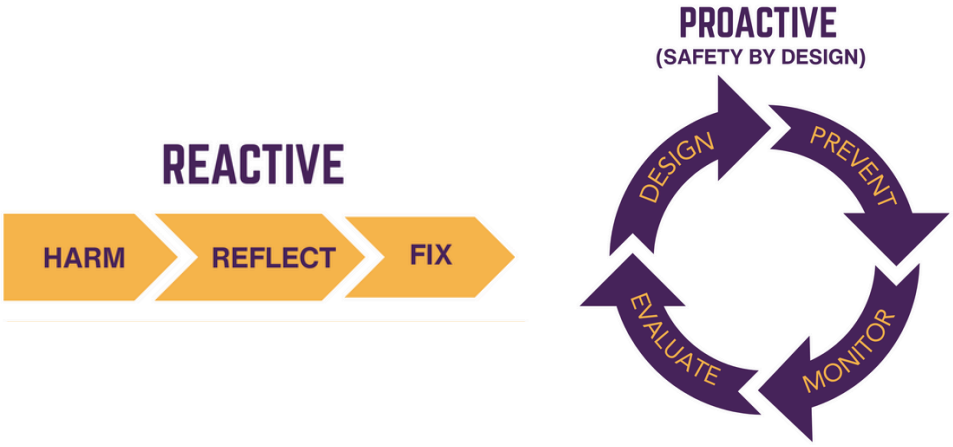
Time Allotted

2 hours

Defining Safety-by-Design

Safety-by-design (SbD) is a proactive approach that puts user safety and rights at the center of the design and development of digital products and services (1). It means finding ways to integrate safety features and considerations from the product or service design and development phase to its launch and use, rather than trying to add safeguards after harm has already occurred.

The primary goal of SbD is to allow digital product and service operators to be able to anticipate, detect, and eliminate potential digital harms before they take place (1). It is about incorporating safety into the culture and leadership of an organization, with a keen emphasis on accountability, transparency, responsibility, and user empowerment. SbD is about preventing and minimizing the potential for digital tool and service misuse from the very beginning (2).



A comparison of traditional reactive approaches and SbD's Proactive approach

Why SbD Matters

Applying safety-by-design as a practice builds trust, operationalizes ethical and duty of care responsibilities, strengthens products' commercial potential, and has positive social impacts.

- **Building Trust & Reputation:** In an environment marked by increasing public and regulatory scrutiny of online products and digital tools, demonstrable commitment to user safety through transparency and proactive measures is essential for building and maintaining user trust and loyalty.
  - **Example:** A mobile money app that clearly communicates its security protocols, implements multi-factor authentication proactively, and transparently reports on potential security incidents is likely to build more user trust compared to an app with unclear security practices or delayed responses to vulnerabilities.

- **Ethical Imperative:** SbD addresses the duty of care that technology providers owe to the individuals who choose to or must use their services, particularly women, girls, and other underserved groups. It provides a concrete framework for fulfilling this ethical obligation by actively working to prevent foreseen or unforeseen harm arising from the use of their products.
  - **Example:** Designing a feature for an online learning platform used by schools to require stringent privacy defaults and clear consent mechanisms for any data sharing fulfills an ethical duty to protect student users who may be required to use the platform for their education.

- **Commercial Benefits:** While investing in safety requires resources, it can also yield tangible business advantages. Digital products and services known for prioritizing safety gain a competitive advantage, attracting the increasing number of users and partners who value safer online environments. Active investment in tools that proactively detect and mitigate harm can also potentially reduce significant long-term operational costs such as incident response management, legal issues, fines, and reputational repair campaigns.
  - **Example:** An e-commerce tool invests heavily in verifying sellers and implementing secure payment systems. While this costs more initially, it gains a reputation for reliability, attracting more customers and premium sellers than competitors plagued by scams, ultimately increasing market share and reducing the costs associated with handling fraud disputes.

Social Impact: The adoption of SbD principles contributes positively to creating healthy digital environments. It does this by supporting the development of safer, more inclusive, and more respectful online spaces. SbD supports more positive online interactions, encourages diverse participation from individuals who might otherwise be silenced, and can help mitigate the detrimental effects of online abuse.

- **Example:** A popular social networking platform uses SbD principles to develop effective, context-aware moderation tools that swiftly address ethnic hate speech and political misinformation. This helps create more constructive dialogue and allows individuals, particularly women and minority groups who might otherwise be targeted, to participate more freely in online discussions.

Understanding the Landscape of Digital Harms & Risks

Digital products and services have a wide array of possible safety risks that can impact users—especially women, girls, and other groups whose needs are often not considered during the design process. These risks can vary based on the unique characteristics of each digital tool or service and can include exposure to harmful content, misuse of personal or sensitive data (such as government-assigned identifiers), and unintended uses of digital products and services that may facilitate violence, grooming, or exploitation. A safety-by-design approach helps identify and protect against these risks by examining user interactions, data collection processes, and the safeguards in place to ensure privacy, security, and user protection.

Examples of risks associated with illustrative specific technologies include (but are not limited to):

Tech Type	Examples of Associated Risks
User-Generated Content (e.g. social media, forums, review sites)	Defamation and reputational harm, hate speech, coordinated online attacks, doxxing, sharing of non-consensual intimate content, exposure to graphic violence, grooming, radicalization, physical harm facilitated by the use of the digital product/service.
Online Marketplaces / Referral Services (e.g. e-commerce platforms, job boards, gig apps)	Fraudulent listings, payment or refund scams, leaks of local and financial information, financial harm/debt due to addictive or gamified design, sale of illegal goods and services, privacy violations and data misuse, malware delivery and malicious scripts.
AI Chatbots and Conversational Agents (e.g. customer support bots, virtual assistants, FAQ bots, in-app helpers)	Data manipulation or falsification, overdependence on automation, lack of autonomy, reduced critical thinking, non-consensual surveillance and tracking, biometric and behavioral data leaks, censorship or profiling of users due to biased algorithms.
Live Interaction and Broadcasting Systems (e.g. video conferencing, livestreaming, webinars, virtual events)	Privacy violations and data misuse, non-consensual surveillance and tracking, exposure to graphic violence, extremist propaganda, and explicit language and images, coordinated online attacks, child sexual abuse material (CSAM).

Tech Type	Examples of Associated Risks
Learning Management Systems (LMS) and EdTech Tools (e.g. online courses, tutoring platforms, classroom tools)	Exposure of children to explicit content (images, video), grooming, hate speech, inaccurate, biased or low-quality information leads to poor or dangerous decisions, malware delivery, malicious scripts, non-consensual profile creation/impersonation.
Gaming and Virtual Worlds (e.g. multiplayer games, immersive environments, metaverse platforms)	Adverse mental and physical impacts due to excessive screen time, algorithmic amplification of harmful or divisive content, biometric/behavioral/lifestyle data leaks, extortion and blackmailing, coordinated online attacks, CSAM, grooming, radicalization.
Content Recommendation Engines (e.g. personalized news feeds, video/music suggestions)	Echo chambers and filter bubbles, exposure to graphic violence, inaccurate, biased or low-quality info leads to poor or dangerous decisions, algorithmic amplification of harmful or divisive content.
Digital Forms and Survey Platforms (e.g. feedback tools, data collection apps, polling tools)	Data leaks, security breaches, hacking, data permanence, difficulty navigating security measures/dark patterns, collection of data without consent, inaccurate data for already marginalized / underserved groups.
Navigation and Location-Based Services (e.g. maps, ride-hailing apps, geofencing tools)	Location data leak, biometric/behavioral/lifestyle data leaks, account hacking, collection of data without consent, fraudulent listings, non-delivery, payment or refund scams, physical harm by another person facilitated by digital product/service, stalking.
Payment and Financial Tech Platforms (e.g. mobile wallets, peer-to-peer payment apps, crypto exchanges)	Financial harm/debt due to addictive or gamified design, financial manipulation into overspending, malware delivery, malicious scripts, non-consensual profile creation/impersonation, tech-based coercive control, sale of unsafe or illegal goods and services.
Tracking and Monitoring Systems (e.g. analytics dashboards, health monitoring, performance tracking)	Account hacking, biased algorithms/AI discrimination against select groups, inaccurate data for already marginalized /underserved groups, leaks, security breaches, hacking of children's data, privacy violations.
Data Collection and Registry Systems (e.g. enrollment tools, case tracking, structured data capture)	Non-consensual profile creation/impersonation, non-consensual surveillance and tracking by external actors, selling of/ monetizing of data without consent, persons with disabilities unable to use product fully or safely.

A CASE STUDY: Kenya's Digital Landscape

While safety-by-design principles are generally applicable across contexts, to be effectively implemented in each specific case, their operationalization must be grounded in country- and user-specific settings, dynamics, and challenges. Not only does this show the importance of the implementation of proactive safety approaches, but it allows digital product operators to be able to carefully design and develop tools that are best suited for the needs of the market.

Kenya is one of the leading nations on the African continent with regards to technology adoption, and more so within the mobile ecosystem. The uptake of 4G and 5G technologies, which allow for faster bandwidths, allowed millions of Kenyans to actively engage within digital spaces (3). With a staggering 53 million people accessing data subscriptions between July and September of 2024, a mobile penetration figure of 131.5%, and smartphone access rate of 72.6%, it is clear that Kenyans are actively engaged in digital spaces, taking advantage of the immense promise brought about by the digital revolution (3). A youthful and increasingly connected society has allowed for access to opportunities in communication, commerce, education, civic participation and social connection.

While there are immense benefits to the Kenyan population, it is becoming more apparent that these same avenues for communication and interactions have also become fertile grounds for various forms of technology-facilitated harm. These are the negative experiences directly encountered by individuals as they interact on digital platforms and use digital products. These experiences are more often than not abusive, deceptive, and exploitative, and disproportionately target women and girls.

Activity 1: Common Technology-Facilitated Harms in Context

Format: Small group discussion and presentation

Participants work in small groups (3 – 4 people) to share, discuss, and note responses to the question below. They are prepared to present briefly to the rest of the group at the end of their conversation.

Discussion question: "Based on your experience or observations using or building digital tools and products in your context, what are some common digital/tech-based/and cyber harms and negative experiences individuals can face?"

Facilitator note: Facilitator elevates and highlights the most reported types of harms after all groups have shared, and mentions that they will be revisited as the training continues.

Defining Harms: Online Violence, Digital Violence, and Technology-Facilitated Gender-Based Violence

SbD in digital products makes all users safer and is good for business. However, without a more deliberate focus on various demographics, it is easy to have “blind spots”. For example, understanding the motivations, skills, experiences, and needs of women and girls, helps decrease gendered digital harms.

The following table is adapted from IREX’s 2024 analysis of Kenya's TFGBV landscape and landscape of redress. It summarizes the terminology used by different stakeholders in Kenya. Consider the similarities or differences within your own context. Are any terms the same?

Term	*Kenyan* Users and References
Technology-Facilitated Gender-Based Violence	International NGOs, multilateral organizations and agencies, organizations implementing foreign assistance programs, civic tech, some private sector, researchers and academia, some advocacy organizations. Umbrella term elevates various digital and online harms to higher level (violence), similarly to GBV.
Online Gender-Based Violence	Local NGOs and advocacy organizations. Umbrella term, more descriptive, also elevates various harms to higher level (violence), similarly to GBV. Use of "online" limits the definition, leaving out aspects covered by "technology-facilitated" definition such as abuse via SMS, and the acknowledgement that violence that starts online often moves offline.
Cyberbullying, Doxxing, Cyberstalking, etc.	Survivor services and support organizations, justice actors, survivors. Tactic focused terms, used to describe specific forms of violence, less resonant in their specificity for comprehensive online safety advocacy and framing.
Online/Cyber Harassment	Government stakeholders, justice actors, law enforcement. Used in police, investigative, and court proceedings; used to describe specific forms of violence. Is both specific and vulnerable to subjective interpretation due to lack of definition. Example: Used in Kenya's Computer Misuse and Cyber Crimes Act.
Non-Consensual Image Sharing, Wrongful Distribution of Obscene or Intimate Images	Justice actors, law enforcement, survivors. Used in the Computer Misuse and Cyber Crimes Act. Used in police, investigative, and court proceedings; used to describe specific forms of violence. Is both specific as named in the Act, and vulnerable to subjective interpretation due to lack of definition.
Hate Speech, Incitement to Violence	Justice actors, law enforcement. Used in the Kenya Information and Communications Act. Used in police, investigative, and court proceedings; used to describe specific forms of violence. Is both specific as named in the Act, and vulnerable to subjective interpretation due to lack of definition.

This shows that there are many ways to talk about these harms. HOW these gendered harms are discussed, though, is not nearly as important as the need for stakeholders to HAVE terms for them.

Pause and Discuss: What terms do you currently use in your work to recognize that women and girls in your client/customer base might have different needs and experiences when they engage with digital devices, tools, and products?

A closer look at the broadest umbrella term – TFGBV – helps understand the gendered nature of the online/digital harms.

Facilitator's Note: *Point out for participants that for simplicity, we will use TFGBV as a term throughout the training. At the same time, they are welcome to refer to gendered digital harms in their own ways - as mentioned, how the abuse and harms are described is not as important as the fact that we are discussing them and focusing on solutions that are important.*

UN Women and the World Health Organization, define TFGBV as "Any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms." TFGBV disproportionately targets women and girls. It takes place both online and offline. There are several key characteristics that highlight its complexity and set it apart from other forms of digital harm, which include:

- **Gender-Based Motivation:** TFGBV is not random. It is rooted in existing gender inequalities, stereotypes and harmful norms within a society. It targets people specifically because of their gender identity and affects them disproportionately due to certain societal structures and norms that have been passed down through the generations.
- **Technology as a Facilitator or Enabler:** TFGBV capitalizes on the unique and ever adapting nature of Information and Communication Technologies (ICTs) and related digital tools as the primary means of enactment, amplification, aggravation, and encouragement of harmful behaviors. This includes a wide range of technologies, including social media platforms, instant messaging applications, email, SMS, online gaming environments, dating apps, location-tracking services, spyware, and AI-driven tools like deepfake generators (4).
- **The Online-Offline Continuum:** TFGBV rarely exists only in the digital spaces. It operates on a continuum where online abuse can and does actively translate into offline spaces in the form of offline threats such as stalking, physical violence, economic harm, or in-person verbal abuse (4).
- **Infliction of Real Harm:** Despite the existence of technologies and tools that act as mediators, the harms caused by TFGBV are evident. They span from psychological impacts, economic and social consequences, political effects, and in some cases, physical harm or an increased risk of physical gender-based violence (4).

Common Forms of TFGBV

Research and analysis including studies conducted by IREX and partner organizations, reveal several common ways in which TFGBV manifests itself. They include:

- **Online/Cyber Harassment:** This involves repeated, unsolicited and unwanted contact, intrusive or offensive messages, public insults, threats of violence (including sexual threats), and often highly sexualized abuse directed towards an individual using digital channels.
- **Cyberbullying:** Characterized by the persistent infliction of psychological or emotional harm through digital technologies, often involving public humiliation, spreading rumors, social exclusion, or coordinated attacks aimed at undermining an individual's self-esteem, reputation, and social standing.
- **Coercive Control:** A pattern of behavior designed to manipulate, intimidate, and control another person. In digital spaces, tech-based coercive control creates fear and oppression of individual's agency and independence. Some controlling behaviors, such as location tracking, sharing passwords, constant messaging, are often framed as acts of caring and love, which normalizes abuse.

- **Image-Based Abuse (including Non-Consensual Intimate Imagery - NCII):** This violation encompasses the creation, manipulation, non-consensual distribution, or threat of distribution of intimate or sexual images and videos of an individual without their explicit consent. This includes cheap- and deepfake video.

- **Cyberstalking:** Involves the persistent use of technologies such as GPS tracking, monitoring of social media and other digital activity (finance, health, etc.) and the use of spyware to monitor, track, harass, and intimidate an individual, creating a persistent sense of fear, intrusion, and loss of privacy, autonomy, and agency.

- **Gendered Disinformation:** This is the intentional creation and spread of false or misleading narratives specifically designed to target women and girls. It often aims to discredit their reputation, undermine their credibility, incite hostility against them, or discourage their participation in public life and the economy, particularly if they are women leaders in politics, journalism, business, or activism.

- **Online Hate Speech:** Consists of abusive, derogatory, or discriminatory language targeting individuals based on their gender, race, ethnicity, religion, sexual orientation, disability, or an intersection of multiple characteristics (5).

- **Doxxing:** The malicious act of openly publicizing an individual's private or personally identifying information such as their home address, phone number, workplace details, and information about family members without their consent. Often the intention is to incite harassment, intimidation, or physical and psychological harm.

- **Other Forms:** The landscape of TFGBV also includes tactics such as impersonation, hacking, and sextortion.

The figure below is a representation of many other forms of TFGBV. It is a good opportunity to revisit the harms that were identified earlier and discuss the overlap or disconnect between individual experiences and aggregated data.

Source: UNFPA (2021)Making All Spaces Safe: Technology-facliliated gender-based vioelence.

TFGBV Prevalence and Impact

The impact of TFGBV cannot be overstated. Its effects are becoming more evident in societies driven by high rates of technology adoption and more so within the mobile internet access ecosystem where a vast majority of people have access to and are increasingly dependent on digital products and services and tools via local internet service providers and access to smartphones.

The impacts of abuse are not siloed into one category of the population or one sector. It affects young and mature uses in urban and rural areas and spans politics, social life, health, entrepreneurship, and other fields of digital engagement.

Example 1: research conducted within Kenya on the prevalence of TFGBV indicates that nearly 90% of students at the tertiary level in the capital, Nairobi, have witnessed TFGBV, while a significant portion of the surveyed group (39%) reported having been directly affected by it. Women constituted about 64.4% of people affected directly by TFGBV, highlighting the gendered nature of this form of digital harm and violence (5).

Example 2: The political arena in Kenya is extremely hostile with research showing that approximately 56% of female political candidates have experienced TFGBV on social media, a significantly higher rate than that of their male counterparts. Additionally, the nature of these attacks differed based on gender, with female candidate being the subject of more sexualized abuse (5,6). These findings correspond with global trends where studies show that most women (about 85%) reported having been witness to violence directed to other women within the digital ecosystem.

The impacts of TFGBV are far reaching and deeply harmful to the individuals it affects. They include:

- **Psychological Impacts:** Common effects include heightened anxiety, chronic stress, clinical depression, symptoms consistent with post-traumatic stress disorder (PTSD), pervasive fear for personal safety, diminished self-esteem and self-worth, and feelings of helplessness or powerlessness.
- **Social Impacts:** Victims often face social stigma, isolation from support networks, damage to personal and professional relationships, and significant reputational harm within their communities or professional fields. More broadly, “normalization” of abusive behaviors in digital spaces feeds into negative trends often abused by those seeking to radicalize the notions of gender equality.
- **Economic Impacts:** TFGBV can lead to tangible economic consequences, including loss of employment, reduced productivity, missed educational or career opportunities, damage to businesses or entrepreneurial ventures, and direct costs associated with mitigating the harm such as legal fees, cybersecurity services and relocation costs.
- **Political and Civic Impacts:** A particularly dangerous consequence is the silencing effect TFGBV has on women's participation in public discourse and civic life. Fear of abuse leads many women to self-censor, withdraw from online digital products and services, avoid expressing opinions, or even abandon careers in fields like journalism or politics. This is a phenomenon often termed the 'chilling effect'.
- **Physical Safety Risks:** The online-offline continuum means that online threats, stalking behaviors, or doxxing can create credible risks of offline confrontation, harassment, or physical violence, forcing victims to alter their routines or take costly security precautions.

Addressing TFGBV

Addressing Technology-Facilitated Gender-Based Violence (TFGBV) requires proactive prevention strategies embedded within the development process of digital tools and services. safety-by-design principles offer a framework for technology developers to anticipate and mitigate harm before products reach users. According to UNFPA's guidance on safe and ethical use of technology to address gender-based violence, implementing safety features during design phases significantly reduces incidents of online gender-based violence compared to reactive approaches (7).

Effective reporting mechanisms represent a critical component of this preventative framework—these accessible, responsive systems increase user confidence while simultaneously decreasing perpetrator behavior through perceived accountability.

A comprehensive safety-by-design approach requires multi-layered reporting mechanisms that include:

- Developing simplified reporting interfaces with local language options.
- Designing reporting systems with features like screenshot capabilities for evidence preservation, anonymous reporting options, and transparent case tracking so users understand what happens after reporting incidents.
- Internal training and practices for roles, functions, and working practices related to the reporting journey and management.
- Incorporating accountability measures like published response time standards and regular transparency reporting on case resolution that recognizes unique threats.
- Integrating simplified guides to obtain users' meaningful consent in reporting mechanisms.
- Implementing trauma-informed responses that prioritize user safety, informed consent, respecting anonymity requests and establishing clear escalation pathways to relevant authorities when necessary.

Examples of effective reporting pathways in Kenya should include platform-level reporting, referral to national institutions such as Communications Authority Kenya Computer Incident Response Team (KE-CIRT/CC), National Computer Cybercrimes Coordination Committee (NC4), international organizations, civil society organizations (FIDA Kenya, CREAW, AMWIK), and dedicated helplines like the National GBV Helpline 1195 and Childline 116.

Reporting builds evidence which drives iterative action, enhances accountability and change.

Activity 2: Testimonial from Survivors

If a suitable local testimonial/ video is available, allow for time for the video/ testimonial to be presented and facilitate a brief reflection.

You may also opt to include the prepared slides on Key Vulnerabilities & Priorities for Action Identified by Survivors of Online Violence on Digital Platforms that are included in the PowerPoint attachment of this curriculum.

Once the activity is completed, take a few moments to provide a summary and reflect on the entire module.

MODULE

2

SbD

CORE PRINCIPLES & FRAMEWORKS

MODULE 2

Core Principles and Frameworks

Learning Objectives

Upon completion of this module, participants should be able to:

1. Identify and explain the 7 core principles of safety-by-design
2. Provide examples of how these principles manifest in existing features
3. Describe prominent SbD frameworks from key organizations (eSafety, TSPA, OFCOM, CCDH, OECD, UNICEF/5Rights)
4. Compare the different emphases and components of these frameworks.

Materials Needed

- Participant handouts
- Markers
- Flipcharts

Time Allotted

2 hours

Safety-By-Design (SbD) Core Principles

At the center of SbD are its core principles and operational frameworks. While there are variations from framework to framework, there are a common set of foundational principles and ideas that underlie most of the approaches, whether they are from regulatory bodies, advocacy groups, or tech industry-led organizations. The principles that form for the foundation for SbD for online products and services are:

1. Proactive Harm Prevention
2. Service Provider Responsibility
3. User Empowerment and Autonomy
4. Transparency and Accountability
5. Holistic Lifecycle Integration
6. Synergy with Privacy and Security
7. Inclusivity and Consideration for All Users

Proactive Harm Prevention

Digital/tech/online service providers should have a forward-looking approach focused on anticipating harms, especially for women and girls, that might be brought about because of the use of their tools and services and finding ways to embed preventative measures at the earliest stages possible in the PDLC (8). It requires service providers to make a shift from reactive approaches to harm to a more proactive approach.

This can be achieved through development and utilization of robust governance frameworks to anticipate, detect, and mitigate harmful behaviors proactively (9). A good starting point is conducting risk assessments early in the product development and throughout the life cycle and proactively engineering out misuse potential (10).

Service Provider Responsibility

Service Provider Responsibility emphasizes that ensuring a harm-free experience for their customers/users is primarily the responsibility of technology service/tools/product providers. Digital products and services are expected to design and implement technological features that prevent abuse and harassment, shifting the bulk of the responsibility from users, who typically bear the brunt of these harms.

This can be achieved by strengthening internal accountability through clear policy enforcement, and active responsiveness to user reports and concerns (20). A good starting point is having a clear duty of care/safeguarding policy that outlines company's commitments to user safety.

User Empowerment and Autonomy

User Empowerment and Autonomy is centered on enabling users of digital tools and products to have substantial control over their experiences while using these products. Users should be provided with accessible and effective tools to manage their interactions through and with the digital product. This would in essence allow them to avoid or mitigate negative experiences proactively (1).

This can be achieved by providing transparent and easy-to-use mechanisms through which users can control their digital presence and interactions, thereby enhancing their independence. Clear, concise, and easy to understand privacy and security information should be provided to the user along with the information on reporting and managing their information should harms occur. A good starting point is to have the strongest privacy and safety features enabled as a default setting ("safety by default") and inform users of what the harms of opting out of them would be should they choose to do so.

Transparency and Accountability

Transparency and Accountability ensures tech service/tools/products providers are open about and accountable for their safety practices and policies. Transparency involves clear communication about how user data is managed, how content moderation (in case of user-generated content platforms) decisions are made, and what actions are taken in response to safety violations. This principle also encourages multi-stakeholder collaboration and feedback mechanisms that allow continuous improvement in safety features and responsiveness to issues brought up by the community of users.

This can be achieved by adopting a practice of tracking disaggregated data (gender, other relevant demographics) on safety concerns and mitigation and progress over time and by regular and transparent reporting on performance related commitments (2). A good starting point is to update product performance indicators and tracking to ensure inclusion of risk monitoring and mitigation and disaggregation of user experiences to understand the nuanced vulnerabilities of different user groups.

Holistic Lifecycle Integration

SbD is not just a one-off checklist item. It is a continuous process embedded throughout the product lifecycle from the ideation and design phases, to development, testing, and deployment and for all subsequent iterations and updates of a product (1).

This can be achieved by designating teams of people or individuals to whom a primary 'championing' responsibility can be assigned and who are responsible for ensuring adherence to SbD at each stage of the cycle. This will contribute to embedding SbD into the culture and processes of the company and ensure that SbD informs priorities, processes, and decision-making frameworks at all levels and for all relevant teams (product, design, engineering, legal, policy, strategy, trust and safety). A good starting point is having clear SbD expectations outlined for each stage of the product development cycle and a person responsible for ensuring they are met.

Synergy with Privacy and Security

Strong privacy and security protections are a crucial part of SbD. Strong security measures protect against unauthorized access and data breaches, while strong privacy protection like data minimization and user control prevent data misuse and empowers users, which contributes to the overall safety of users (11).

This can be achieved by instituting mandatory privacy and security reviews that account for different types of potential users and their needs and vulnerabilities during safety assessments of all products before they are launched (8). A good starting point is making sure that these checks are tested in a pre-launch "abusability test".

Inclusivity and Consideration for All Users

For SbD to be effective, service providers must design their tools, safety features, and prevention and mitigation strategies with an awareness of the needs, capabilities and vulnerabilities of the entire user community, not just the most dominant/represented segment. This includes deliberate assessment of potential harms faced by categories of users such as legal minors (children under 18), users with accessibility needs, new digital arrivals and individuals of varying levels of digital literacy, and those who face disproportionate risks online and in digital spaces - women and girls (12).

This can be achieved by integrating the full scope of potential users/customers in market research and exploring the needs, motivations, vulnerabilities, and risks associated with categories of potential users, and subsequently designing features that make digital offerings more attractive and safer for them. A good starting point is to employ "digital personas" – composite profiles/typology of users, to inform the design process.

Practical exercise – let's explore some of the existing SbD features and identify what principles they address.

ACTIVITY 3 - Demo Session: Practical Examples of SbD Features

Format: Facilitator-led demonstration and plenary discussion. Use the table of features and corresponding principles below to guide the activity.

Facilitator demonstrates each selected feature (e.g., using screenshots of Instagram's privacy settings, WhatsApp's blocking mechanism, Twitter/X's reporting flow, Bumble's content blurring).

For each feature, the facilitator explicitly identifies:

- The **safety-by-design feature** being shown (e.g., Granular Privacy Controls, User Blocking, Clear Reporting Mechanism, Content Filtering).
- The primary **SbD principle(s)** it embodies (e.g., User Empowerment, Provider Responsibility, Proactive Prevention, Transparency).
- The specific **safety goal** or harm it aims to address (e.g., preventing unwanted contact, enabling users to flag abuse, protecting users from unsolicited graphic content).

SbD Feature	Corresponding Principles	Example
Safe Defaults	Proactive Harm Prevention, Inclusivity, Provider Responsibility	Instagram automatically setting accounts for users under 16 to private. Protects younger users by default, requiring them to consciously opt-in to public visibility, reducing unsolicited contact or exposure.
Customizable Privacy Settings	User Empowerment & Autonomy, Transparency	Facebook's granular controls allowing users to decide who sees their posts (Public, Friends, Custom lists), tags, or personal information. Gives users direct control over their information visibility and interaction surface.
End-to-End Encryption (E2EE)	Synergy with Privacy & Security, User Empowerment	WhatsApp ensuring only the sender and recipient can read messages. Protects message content from interception by third parties or even the tool provider, crucial for confidential communication and preventing surveillance.
Two-Factor Authentication (2FA)	Synergy with Privacy & Security, User Empowerment	Google Accounts or online banking apps requiring a code from an SMS or authenticator app in addition to a password. Adds a layer of security to prevent unauthorized account access, protecting personal data and control.

SbD Feature	Corresponding Principles	Example
Granular Permission Controls	User Empowerment & Autonomy, Synergy with Privacy & Security	Android/iOS allowing users to grant/deny app access to specific data/features (Location, Camera, Microphone, Contacts). Empowers users to limit data exposure and prevent apps from overreaching, reducing privacy risks.
Clear Reporting Mechanisms	User Empowerment, Provider Responsibility, Transparency/Accountability	Twitter/X's feature to report specific posts or profiles for violations (harassment, hate speech) with clear categories. Provides users a direct channel to flag harm, enabling action and accountability.
User Blocking & Muting Tools	User Empowerment & Autonomy	Most social media platforms allowing users to block others (preventing all interaction) or mute them (hiding their content without notification). Essential tools giving users immediate control to stop unwanted interactions or content.
Content Filtering / Sensitivity Screens	Proactive Harm Prevention, User Empowerment	Instagram or Bumble blurring potential nude images in Direct Messages (DMs) with a warning, allowing users to choose whether to view. Protects users from unsolicited graphic content while giving them control over exposure.
Automated Content Moderation (AI/ML)	Proactive Harm Prevention, Provider Responsibility	YouTube using AI to proactively detect and remove content violating policies (e.g., CSEM, hate speech) often before user reports. Scalable method to reduce harm exposure, though requires oversight for accuracy/bias.
Hashing for Known Harmful Content	Proactive Harm Prevention, Provider Responsibility	Tools like StopNCII.org allow users to hash intimate images; digital service providers use these hashes to block re-uploads without seeing the image. Prevents the spread of known illegal/harmful content like NCII and CSEM across providers.
Identity / Age Verification Mechanisms	Proactive Harm Prevention, Provider Responsibility, Inclusivity	Dating apps requiring profile verification (e.g., photo matching) or digital products using methods to verify age for age-gated content/features. Can enhance accountability and protect minors but requires careful design re: privacy/bias.
Transparency Reports	Transparency & Accountability, Provider Responsibility	Meta or Google publishing regular reports detailing content removed, government requests, and enforcement actions. Provides public insight into safety efforts, trends, and adherence to policies, fostering accountability.

SbD Feature	Corresponding Principles	Example
Inclusive Design (applied to Safety)	Inclusivity & Consideration for All Users	Designing reporting flows to be navigable via screen readers or providing safety information in multiple languages. Ensures safety features and information are accessible and usable by people with diverse needs/abilities.
Data Anonymization / Privacy Enhancing Tech	Synergy with Privacy & Security	Apple using Differential Privacy to analyze user trends without accessing individual raw data. Techniques allowing digital products and services to gain insights or train models while minimizing exposure of identifiable user information.
User Consent Management	User Empowerment, Transparency, Synergy with Privacy & Security	Websites/apps presenting clear cookie banners or data usage consent prompts aligned with GDPR/DPA principles. Ensures users have informed control over how their data is collected and used, a cornerstone of data protection.
Automated Fraud Detection	Proactive Harm Prevention, Synergy with Privacy & Security	M-Pesa or PayPal systems analysing transactions to flag and block potentially fraudulent activity in real-time. Protects users from financial loss and economic abuse, a relevant aspect of online safety.

Prominent SbD Frameworks

The evolution of SbD has been driven by the innovative and competitive tech sector – as you see many of the examples above come from very successful companies and by human/consumer rights and protections agencies and organizations with mandates to protect individuals. These entities have developed several complimentary SbD frameworks. The next section of the training will review these approaches and discuss the subtle differences that they contain.

Australian eSafety Commissioner

Australia's independent government agency dedicated to online safety. Its mandate is to safeguard Australians from online harms and promote safer online experiences through education, regulation, research, and reporting schemes for issues like cyberbullying, image-based abuse, and illegal content.

- Website: https://www.esafety.gov.au/
- SbD Resource: Safety-by-Design Initiative & Tools
- TFGBV Resource: Technology, Gendered Violence and Safety-by-Design: An industry guide

Australia's e-safety Commissioner's Safety-by Design Framework is one of the most influential and applicable of the existing frameworks. It was developed by the eSafety Commissioner in Australia as a means of providing practical guidance, assessment tools and practical steps for use by industry players with the aim being to make SbD as realistic and achievable for companies of all sizes. At its heart are three core guiding principles: Service Provider Responsibility, Empowerment and Autonomy, and Transparency and Accountability (10). The tools and resources provided by the eSafety commissioner were developed through extensive consultation with key stakeholders including industry players, NGOs, and users, making it one of the most versatile frameworks that exist today.

Trust and Safety Professionals Association (TSPA)

A non-partisan, non-profit membership association supporting the global community of professionals who develop and enforce principles defining acceptable online behaviour and content. TSPA aims to build a community of practice and provide resources for trust and safety work.

- Website: https://www.tspa.org/
- SbD Resource: TSPA Safety-by-Design Curriculum

The TSPA provides a curriculum on trust and safety which highlights five key principles. These include Holistic, Proactive, Empowering, Synergistic (with privacy/security), and Transparency with regards to building solutions that are safe for users. It provides a good amount of detail on the process of implementing these principles (risk assessment, red teaming), technical interventions (proactive detection, user controls), and instruments to measure how effective implementation is. The aim here is to assist tech organizations in operationalizing SbD as efficiently as possible (8).

UK Government: Office of Communications (Ofcom)/ National Cyber Security Centre (NCSC)

Ofcom is the UK's communications regulator, responsible for enforcing the Online Safety Act, which mandates safety measures, particularly for children. The NCSC is the UK's technical authority on cyber security, providing guidance on secure system design.

- Website (Ofcom): https://www.ofcom.org.uk/
- Website (NCSC): https://www.ncsc.gov.uk/
- Ofcom Resource: Ofcom calls on tech firms to make online world safer for women and girls
- NCSC Resource: NCSC Secure Design Principles

Developed in response to the Online Safety Act , the Ofcom approach puts emphasis on service provider responsibility, focused primarily on child protection (12). Section 13 of the act seeks to ensure that services that are regulated are Safe by Design and designed and operated in a way that higher standards of protection are provided for children more than adults. It also ensures that user rights to free expression and privacy are protected, and transparency and accountability are guaranteed relative to these services (12). This framework also provides practical tools and guidance to ensure that women and girls remain safe online through examples of abusability testing technologies to prevent intimate image abuse, easier account controls, visibility setting and others (13).

STAR Framework (Center for Countering Digital Hate)

CCDH is an international non-profit organization that works to disrupt the architecture of online hate and disinformation through research, public campaigns, and policy advocacy, aiming to hold social media companies accountable.

- Website: https://counterhate.com/
- SbD Resource: STAR Framework (Safety-by-Design, Transparency, Accountability, Responsibility)

The STAR Framework was developed in partnership with regulators, legislators, civil society and academia from the UK, US, EU, Canada, Australia, and New Zealand who came together to reflect on the current state of digital harm and to compare notes on what different global players in the space were doing (14). It emphasizes that tackling online hate and misinformation requires a multi-faceted approach built on four interconnected pillars:

1. Safety-by-Design (S): This is positioned as the foundational approach. It mandates that digital products and services proactively design their systems and features to prevent harm from occurring in the first place, rather than relying solely on reactive measures after harm emerges. It requires embedding safety considerations throughout the product lifecycle.

2. Transparency (T): This pillar calls for openness regarding key digital products and service operations, including the functioning of algorithms, the processes for rule enforcement, and the economic incentives driving design choices (such as making unsubscribe options difficult to find or specific advertising economics).

3. Accountability (A): This focuses on ensuring digital products and services are answerable to independent and democratic oversight bodies, moving beyond self-regulation.

4. Responsibility (R): This highlights the duty of companies and their senior executives not only to their users but also to the overall health and integrity of the information ecosystem.

This framework looks at SbD within the broader context of governance, focusing on the need for proactive safety standards, similar to those that exist in the consumer sectors (14).

Organization for Economic Co-operation and Development (OECD)

An international organization focused on developing better policies globally. In the digital sphere, it provides analysis and develops standards and recommendations related to digital security, privacy, and specifically, child online safety, promoting responsible digital transformation.

- Website: https://www.oecd.org/
- Child Safety Resource: OECD Recommendation on Children in the Digital Environment

The OECD takes into account and combines learning from Security by Design and Privacy by Design in the development of key insights into Safety-by-Design specifically for children (15). To achieve SbD, they recommend:

1. **Proactive approaches:** Systematically identifying and addressing potential vulnerabilities before they can be exploited to harm children (e.g., anticipating misuse scenarios specific to child users).

2. **Ongoing and dynamic risk management:** Continuously assessing and adapting to evolving risks related to children's safety online, considering changes in technology, usage patterns, and the developmental stages of child users, rather than relying on static, one-off assessments.

3. **User-centric design (Child-centric design):** Designing services and safety features specifically with children's best interests, developmental needs, rights, cognitive abilities, and diverse experiences at the core, ensuring usability and appropriateness for different age groups.

4. **Accountability:** Establishing clear responsibility within the organization for implementing and overseeing child safety measures, including mechanisms for redress and compliance verification.

5. **Transparency:** Clearly communicating to children (in age-appropriate ways) and their caregivers about risks, safety features, data practices, and how the service operates.

At the core, they provide principles that include upholding children's best interests, empowerment and resilience, as well as a balanced approach to digital consumption and use, age-appropriateness, inclusion and shared responsibility amongst stakeholders (16). It also provides guidelines for the providers of digital services on SbD, data protection and governance.

UNICEF and Child Rights Perspectives

UNICEF is the United Nations agency dedicated to protecting the rights of every child. 5Rights Foundation is an international organization advocating for children's rights and safety in the digital environment, pushing for systemic change by design. Together, they champion the 'Child Rights by Design' (CRbD) approach.

- Website (UNICEF):https://www.unicef.org/
- Website (5Rights):https://5rightsfoundation.com/
- CRbD Resource:UNICEF: The children's rights-by-design standard for data use by tech companies

UNICEF and organizations like 5Rights advocate for a 'Child Rights by Design' approach which integrate the principles from the UN Convention on the Rights of the Child (CRC), into the product life cycle for digital services, and especially those used by children (17). It introduces specific methodologies and techniques which include strong age verification methods, effective default privacy settings, protection from commercial exploitation and providing avenues for redress (17).

The following table provides a comparative overview of the different frameworks and their core principles

Principle Theme	eSafety Commissioner (Australia)	TSPA	UK Govt/Ofcom	OECD (Children Focus)	UNICEF/CRbD Focus
Proactivity/ Prevention	Assess/address harms upfront; engineer out misuse; **prevent harms** before occurring	Target prevention; **early risk assessment;** red teaming	Preventative steps; **reduce exposure**; tackle harms proactively (esp. child sexual exploitation and abuse (CSEA), grooming)	**Child Safety-by -Design;** precautionary approach; prevent access to harmful content	Preventative steps; evaluate known/ anticipated harms in design; **proactive detection**
Provider Responsibility	**Service provider responsibility;** burden not solely on user; accountable teams	**Holistic;** Applies organizationally; **Synergistic** - Commit to secure practices	Users not left to manage own safety; provider duties under OSA; **accountability for compliance** [42]	**Shared responsibility;** digital service providers have key role; governance & accountability; respect guidelines	Service provider responsibility; shared responsibility; **accountable for identifying/ addressing harms**

Principle Theme	eSafety Commissioner (Australia)	TSPA	UK Govt/Ofcom	OECD (Children Focus)	UNICEF/CRbD Focus
User Empowerment /Autonomy	User empowerment & autonomy; **dignity;** user controls; default safe settings	Empowering; user controls; **meaningful engagement**; shape own experiences	Empower users for safer decisions; user controls; tools to manage safety; **default high settings** (esp. kids)	**Empowerment & resilience;** support users; provide controls; uphold rights (expression, participation)	User empowerment & autonomy; align with best interests; consult users; **accessible features**
Transparency/ Accountability	Transparency & accountability; open communication ; **accessible policies**; publish metrics	Transparency; openness enhances safety; **transparency reports**; clear rules	Transparency reports; clear terms and conditions; **accountability via regulator** (Ofcom)	Information provision & transparency; **clear, age-appropriate information**; accountability mechanisms	Transparency & accountability; **share enforcement data**; effectiveness of features; share innovations
Inclusivity/ Vulnerability	Consider diverse users; **mitigate risks for distinct groups;** accessibility	**User consultation;** consider identity/ circumstances	**Consider all user types** (protected characteristics , kids, accessibility, literacy)	**Appropriateness & inclusion;** age-appropriate; account for diverse needs; avoid discrimination/ bias	Consult diverse/at-risk groups (inc. children); **accessible features;** protect rights
Lifecycle Integration	**Incorporate throughout design,** provision, assessment	Holistic - Safety across product lifecycle; **organizational integration**	Design and operate safely; **risk assessments throughout**	Consider **safety in design, development, deployment, operation**	**Embed child rights in entire product life cycle and provision of service**
Rights-Based Approach	Preserve fundamental **consumer/ human rights**	(Implied through **user empowerment/ privacy)**	(Implied through legal duties, **focus on illegal harms)**	Child's best interests; protect/respect rights; **respect human rights**	**Child Rights by Design;** based on CRC; best interests; protect all rights

Table 1: Comparison of SBD Principles Across Key Frameworks

Module 2 Summary & Plenary Exercise: Principles in Practice

Summary: Facilitator briefly recaps Module 2, highlighting the 7 core SbD principles and the introduction of various organizational frameworks (eSafety, TSPA, Ofcom, CCDH, OECD, UNICEF/CRbD). Emphasize that while frameworks differ slightly in focus, they share fundamental goals.

Plenary Discussion:

Facilitator poses the following questions to the group, encouraging broad participation:

- "Reflecting on the different frameworks (eSafety, TSPA, UK, OECD, etc.), what stood out to you as the most important common theme or shared emphasis regarding safety-by-design?"
- "Considering the 7 core SbD principles we discussed (Proactivity, Responsibility, Empowerment, Transparency, etc.), which one do you anticipate might be the biggest challenge to implement consistently within your own work context or a typical tech development cycle? Why?"
- "Conversely, which principle seems the most readily achievable or might offer the quickest wins for improving safety in your products/tools?"

Capture Insights:

Facilitator captures key responses and common threads from the discussion on the flip chart or whiteboard.

Facilitator Role:

Guide the discussion, ensure multiple voices are heard, keep track of time, synthesize the key reflections shared by the group, and provide a concluding bridge statement emphasizing that Module 3 will delve into how to operationalize these principles and overcome implementation challenges within the product development lifecycle.

MODULE

3

FROM PRINCIPLES TO ACTION:

Operationalizing SbD

MODULE 3

From Principles to Action: Operationalizing SbD

Learning Objectives

Upon completion of this module, participants should be able to:

1. Describe key processes for operationalizing SbD (Risk Assessment, Threat Modelling, Privacy/Security Reviews, Abusability Testing, User Consultation)
2. Apply a structured process for identifying potential harms, assessing impact, and developing mitigation controls.
3. Understand the basics of Threat Modelling (STRIDE) from an attacker's perspective.
4. Explain the importance of integrated privacy/security reviews and abusability testing.
5. Map SbD activities onto the stages of the Product Development Lifecycle (PDLC).
6. Apply risk identification specifically to TFGBV harms.
7. Understand the relevance of regulations to SbD.

Materials Needed

- Flipcharts
- Markers
- Handouts (including Risk Matrix Template, PDLC Worksheet Annex)

Time Allotted

2.5 hours

Activity 4: Ideation of Digital Products and Mapping Digital Personas and User Needs

Instructions for Participants:

Phase 1: Group Formation & Tool Ideation (20 minutes)

1. **Form Groups:** Organize yourselves into groups of approximately 6 members
2. **Brainstorm & Select Idea:**
   - As a group, brainstorm potential ideas for a simple digital product or tool.
   - Consider needs or opportunities within the local context (e.g., community support, local commerce, information access).
   - Select ONE idea your group will focus on for this activity.
   - **Output:** Clearly write down your chosen idea and a one-sentence description of its core function.

   Example: "Idea: A mobile app for verified local errand runners in urban areas of Kenya."

Phase 2: Digital Persona Development & Needs Mapping (10 minutes)

1. Identify Diverse Personas:
   - For your group's chosen idea, identify at least three to four distinct digital personas who might use or be significantly impacted by it.
   - Strive for diversity in your personas, considering factors such as:
     - Age, gender, occupation.
     - Location (urban/peri-urban/rural).
     - Digital literacy and typical technology access (e.g., basic smartphone user vs. tech-savvy individual).
     - Primary language.
   - **Include at least one persona representing a group that might be particularly vulnerable to online harms in the local context** (e.g., a young woman, an elderly person, an individual with low digital literacy, an activist).

Phase 3: Group Review & Preparation for Sharing (10 minutes)

1. **Review Personas:** As a group, briefly review the personas you've developed.
   - Do they represent a good range of potential users and impacted individuals?
   - Are the safety needs and vulnerabilities clearly articulated, especially concerning the local context and TFGBV?
2. **Prepare for Share-Out (if applicable):** Briefly describe the idea that you came up with for the digital product and the personas that you have developed (each group)

"How to" of Safety-by-Design - Overview

Operationalizing SbD principles requires a combination of structured processes, specific technological interventions, and deliberate design choices integrated into the product development workflow. Key processes that we will discuss in more detail in this session include:

1. Risk identification and prevention and mitigation planning.
2. Threat modelling.
3. Integrated privacy and security reviews.
4. Safety oriented abusability testing/ red teaming.
5. User consultation and co-design.

Risk Identification and Prevention and Mitigation Planning

With the recognition that safety must be considered and applied to every stage of a product's life cycle, there must be a systematic approach to identifying potential issues and harms and designing potential solutions before the product is deployed.

This process can be broken down into 5 key steps:

1. **Identify Potential Harms/Hazards:** This involves asking critical questions:

   a. What could go wrong with this product or feature?
   b. How might it be misused, even unintentionally?

This requires broad thinking and perspective taking, considering technical vulnerabilities, user behaviors, and potential policy gaps, as well as contemplating who might cause harm (10).

2. **Identify Who Might Be Harmed and How:** This step considers the potential vulnerabilities of users, different user segments, bystanders, and the product's reputation. It also involves specifying the negative consequences, such as emotional distress or loss of privacy (18).

3. **Evaluate Potential Impact & Prioritize Action:** For each identified harm, an assessment of its likelihood (how probable is it?) and severity (how damaging would the consequences be?) is conducted. This evaluation, often visualized using a simple matrix, helps prioritize which harms demand the most urgent attention and resource allocation.

4. **Develop and Implement Prevention/Mitigation Controls:** This involves brainstorming and selecting specific actions. Can the harm be entirely prevented through design changes (UI/UX e.g. making reporting buttons more prominent or adding warning labels before users share sensitive information), policy enforcement, or technical means (altering underlying code and implementing algorithms e.g. deploying content moderation algorithms, implementing end-to-end encryption)? If complete prevention is not feasible, how can its likelihood or severity be reduced? This involves designing reporting tools, implementing moderation systems, and issuing clear warning (10).

5. **Plan for Review and Iterate:** While the actual review of effectiveness happens after deployment, planning for how you will monitor and evaluate your controls is important. This involves thinking about how you will know if the implemented controls are functioning as intended. This step then extends into the post-launch phase where you actively monitor performance by analyzing user reports, feedback, incident data, and other relevant metrics. This ongoing monitoring and analysis fuels a continuous cycle of review and improvement, allowing safety measures to be refined and updated based on real-world performance and evolving threats.

Facilitator's Note: Emphasize this point to participants. While Steps 1-4 heavily focus on pre-deployment design and planning, Step 5 highlights that safety-by-design is not a one-time task. It's important to plan during the design phase how effectiveness will be measured post-launch. This step bridges the planning phase with the ongoing maintenance and iteration phases of the product lifecycle, reinforcing that SbD requires continuous analysis, monitoring, and improvement. Introduce the eSafety risk assessment tool as well (https://www.esafety.gov.au/industry/safety-by-design/assessment-tools).

The risk identification, prevention and mitigation planning process

Activity 5: Guided Risk Identification & Mitigation Simulation

Recall Your Product or Tool Idea:

- Briefly revisit your group's digital product concept and its intended users.

Brainstorm Potential Risks (10 mins):

- For your product, identify at least 3-5 potential safety risks or ways it could be misused.
- *Consider:* Harms to users, TFGBV, data privacy issues, scams, misinformation, vulnerable groups.

Document in Risk Register (20 mins):

- Choose at least two (2) significant risks from your brainstormed list.
- For each chosen risk, use your Risk Register handout/sheet to document:
  - The Risk Description, Potential Victims, and Nature of Harm.
  - Initial Likelihood & Severity scores (use Risk Definitions sheet for scales).
  - Note the auto-calculated Initial Risk Level.
  - Detail your "Proposed Prevention / Mitigation Controls" (think design, policy, tech solutions applying SbD principles).
  - Re-assess Likelihood & Severity after considering your controls, then determine and record the "Residual Risk Level."
- (Follow the detailed instructions on your Risk Register template for guidance on each column).

Prepare to Share (5 mins):

- If time allows for a share-out, select one risk your group analyzed. Be ready to briefly present its initial assessment, your key mitigation ideas, and the resulting residual risk.

Facilitator's Note: Ensure each group has their product concept clear before brainstorming. Encourage them to be specific about potential harms relevant to the local context and to think critically about practical mitigation controls applying SbD principles. Guide them to use the Risk Register template methodically, focusing on how their proposed controls reduce the likelihood and/or severity of the identified risks. If a share-out is planned, select groups to present diverse risks and mitigation strategies.)

Threat Modelling

Threat modelling is another structured approach that allows service providers and development teams to be able to identify, quantify, and address security risks associated with a product or tool that they are developing (19). Threat modelling looks at a solution from the perspective of an attacker instead of a user or someone working to defend a system, helping in increasing product security.

Threat modelling can be broken down into four high level steps (adapted from the OWASP Threat Modelling framework) which include:

1. **Determine the scope of your work:** Before identifying threats, it is important to have a clear understanding of the system, features or processes and its boundaries. To achieve this, the following actions should be taken:
   - Create/ review system diagrams: These may include data flow diagrams (how data moves), architectural diagrams (components and connections) or process flow diagrams. These diagrams help everyone involved understand the components, their interactions, where data flows to and from and what makes up the systems boundaries (what you need to protect).
   - Identify entry points: Determine all the ways external actors (users, other systems, potential attackers) can interact with the system.
   - Map access rights and trust boundaries: Understand who (different types of users like guests, registered users, administrators) or what (other internal/external systems) has access to different parts of the system and what level of permission or trust they have.
   - Understand use cases/user stories: Define how the system or feature is intended to be used legitimately. Reviewing user stories or typical workflows helps establish a baseline of expected behavior. Understanding the intended use is crucial for identifying potential misuse scenarios or how legitimate functions could be abused by an attacker.

2. **Determine threats:** At the center of the process of identifying threats is the categorization method called STRIDE. We are going to try using the STRIDE framework as a helpful tool for identifying vulnerabilities. The word STRIDE itself can serve as a useful mnemonic for different types of threats:
   - Spoofing (impersonation)
   - Tampering (data manipulation)
   - Repudiation (denying actions)
   - Information Disclosure (privacy breaches, doxxing)
   - Denial of Service (disruption)
   - Elevation of Privilege (unauthorized access)

3. **Determine countermeasures and mitigations:** A vulnerability once identified can be dealt with by deploying a countermeasure. These countermeasures can be identified using threat-countermeasure mapping lists that need to be continuously updated.

Sample Threat-Countermeasure Mapping List

Threat	Countermeasure
Spoofing identity	Use strong authentication
Tampering with data	Use digital signatures, protect data in transit
Repudiation	Log actions, use secure time stamps
Information disclosure	Encrypt sensitive data, implement access controls
Denial of service	Perform input validation, manage resource quotas
Elevation of privilege	Enforce least privilege, apply security patches

4. Assess your work: This would be the final step in ensuring that the countermeasures deployed are actually mitigating the identified threats.

Integrated Privacy and Security Reviews

Conducting safety assessments in most if not all cases is done alongside or integrated with privacy impact assessments (PIAs) and security reviews. This helps in ensuring that solutions are compliant with regulations around data protection such as Kenya's Data Protection Act (DPA 2019 Kenya) and the European Union's General Data Protection Regulation (GDPR- EU). It also ensures that security vulnerabilities that could compromise user safety are addressed beforehand to minimize the chances of harms taking place.

While distinct disciplines, reviewing them together ensures that safety mitigations don't inadvertently create privacy risks, and security measures adequately support safety objectives. Key activities typically included in integrated privacy and security reviews:

- **Data Minimization Assessment:** Verifying that the feature collects only the minimum personal data necessary for its specific, legitimate purpose. (Question: Is all collected data truly needed? Can the goal be achieved with less data?)
- **Purpose Specification & Consent Check:** Ensuring the reasons for collecting and processing data are clearly defined, communicated to users, and that appropriate, informed consent is obtained, particularly for sensitive data. (Question: Is it clear to users why this data is needed? Is consent freely given, specific, informed, and unambiguous?)
- **Security Vulnerability Assessment:** Identifying technical weaknesses in the feature's design or code that could be exploited. This might involve security code reviews, checks for common vulnerabilities (like those in the OWASP Top 10), or basic penetration testing to find flaws that could lead to unauthorized access, data breaches, account takeover, or other safety incidents.
- **Change Impact Assessment:** Evaluating how modifications to existing features or systems (including updates, new integrations, or policy changes) affect user privacy, security, and safety. This helps ensure that previously mitigated risks are not reintroduced and that new risks are captured before release.
- **Access Control Verification:** Reviewing who (internal staff, system roles, types of users) has permission to access user data generated or used by the feature, ensuring the principle of least privilege is applied and user-facing controls (like blocking) function correctly at a technical level.
- **Data Handling, Storage & Encryption Review:** Checking that personal data is handled securely throughout its lifecycle – including encryption (both in transit and at rest where appropriate), secure storage practices, and adherence to defined data retention and deletion policies.
- **Internal and External Assessments & Audits:** Conducting regular reviews by both internal teams and external, independent experts to validate that privacy, security, and safety commitments are being met. These audits provide accountability, highlight blind spots, and ensure compliance with evolving standards and regulations.
- **Third-Party Risk Assessment:** Evaluating the privacy and security practices of any third-party services integrated with the feature (e.g., analytics providers, external APIs) to ensure they don't introduce unacceptable risks.
- **Compliance Check (DPA/GDPR etc.):** Explicitly reviewing the feature against the requirements of relevant data protection laws, including rights management (access, correction, deletion requests) and breach notification procedures. This often involves completing or updating a formal PIA, especially for high-risk processing activities.

Safety- Oriented Red Teaming/ Abusability Testing

An often-overlooked aspect of safety-by-design is considering how features, even those designed with good intentions, could potentially be weaponized or misused by malicious actors (9). This involves abusability testing, deliberately thinking through potential misuse scenarios during the design phase to build in preventative measures or necessary mitigations (13). It involves exercises, often hands-on, where external or internal experts act as adversaries to proactively identify vulnerabilities, potential avenues for abuse and ways that a tool, product or feature can be weaponized or misused before launch. Crucially, for safety-by-design, this testing should be conducted proactively during the design and development phases (on mockups, prototypes, or pre-release versions) to identify and address flaws before launch (13). While testing can also occur reactively on live products, the SbD emphasis is on pre-launch prevention. Combined with threat modelling, it provides a broad view of threat exposure and can be used to help build organizational buy-in for necessary safety mitigations.

A few examples illustrating this concept:

- **Location Sharing:** Intended for coordinating meetups but can be misused for stalking or doxxing (20). Potential Mitigations: Offer granular controls (share precise vs. general location), temporary sharing options, clear indicators when location is being shared, default to less precision.
- **Content Editing:** Intended for correcting errors but can be misused to spread disinformation by altering context after initial engagement or to retroactively insert harmful content. Potential Mitigations: Implement visible edit histories, clear labelling of edited content, time limits on editing.
- **Anonymity Features:** Intended to protect privacy or enable sensitive disclosures but can be misused to shield harassers or spread abuse with impunity (20). Potential Mitigations: Carefully balance anonymity with internal accountability mechanisms (e.g., internal logging, stricter enforcement for anonymous accounts engaging in abuse), offer verified identity options.
- **Direct Messaging:** Intended for private conversation, but frequently misused for unsolicited harassment, scams, threats, or grooming (20). Potential Mitigations: Implement message request systems from non-contacts, utilize content scanning/filtering for harmful content, ensure robust in-chat reporting and blocking.

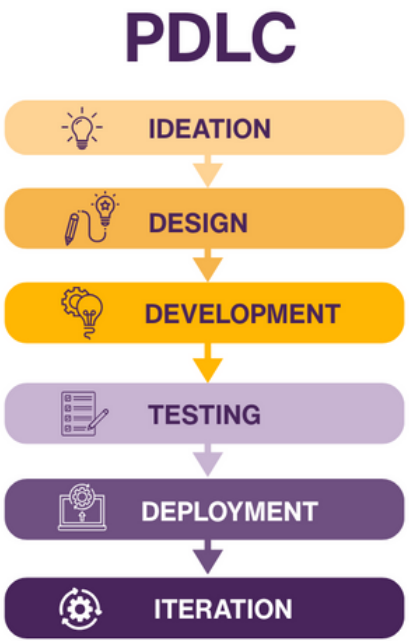
Safety-Oriented Red Teaming/ Abusability Testing

To say that one has understood a harm or a threat, without consulting with and integrating the people it affects directly from the very beginning, would be a falsehood. Engagement with a diverse user community representative of all groups who might use the product/tool/service, including, illustratively, women, civil society, subject matter experts, persons with disabilities, parents and caregivers of young children, rural users with less access to digital skills, linguistic minorities, etc. is key as it helps organizations to be able to understand risks and harms from different perspectives, allowing them to design appropriate and effective safety features and controls.

Mapping SbD Activities to the Product Development Life Cycle

The Product Development Life Cycle

To effectively integrate safety-by-design, it is necessary to identify specific safety-related activities and considerations important to each stage of the Product Development Lifecycle (PDLC), with a clear emphasis on actions aimed at prevention and mitigation. Visualizing a typical PDLC often involves stages such as Ideation/Concept, Design/Planning, Development/Implementation, Testing/Quality Assurance, Launch/Deployment, and ongoing Maintenance/Monitoring/Iteration. While the specifics might vary between organizations, the core progression remains similar.



The Product Development Life Cycle

Integrating SbD means incorporating safety checks, considerations and requirements at each state of the PDLC:

1. Ideation: Identify Harms, Define Safety Goals, Consider Digital Personas of all potential users.
2. Design: Conduct Threat Modelling, Apply Privacy by Design, Design Mitigation Features (e.g., Reporting), Assess Risks.
3. Develop: Implement Secure Code, Build-in Privacy Defaults, Ensure Feature Robustness.
4. Test: Perform Security & Abusability Testing, Verify Safety Features with Diverse Users, Mitigate Risks Before Release.
5. Launch: Communicate Safety Expectations, Implement Safe Defaults.
6. Maintain: Monitor for Harm, Analyse Reports, Respond to Incidents, Iterate for Prevention.

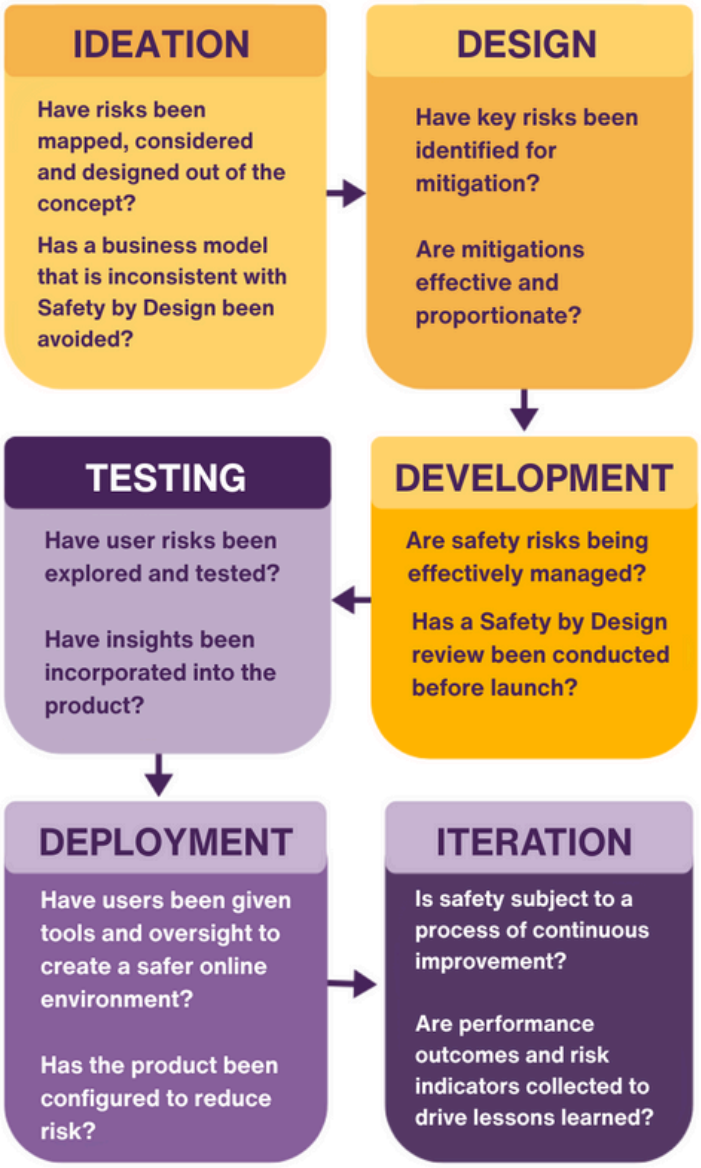
Mapping SbD Activities By Stage

This segment walks through a typical PDLC, highlighting specific SbD-related questions, activities, and outputs relevant to each stage. It demonstrates that safety is not a separate track but an integrated lens.

#	Phase	SbD Focus	Activities	Output
1	Ideation / Concept Phase	Early identification of potential high-level safety risks inherent in the core concept. Alignment with organizational safety values.	• Initial brainstorming on potential misuse scenarios. • Considering vulnerabilities of all potential user groups, not just the dominant one. • Asking: "Could this core idea be easily weaponized? Does it align with our safety principles?"	High-level safety considerations documented alongside the initial concept brief.
2	Design	Translating safety principles into concrete design choices (UI/UX), feature specifications, and policy requirements. Detailed risk assessment. Threat modelling to identify potential risks prior to development.	• Conduct formal Safety Risk Assessments. • Design specific mitigation features (e.g., robust reporting mechanisms, user-friendly blocking tools, content filtering options, clear warnings). • Make deliberate User Interface (UI) and User Experience (UX) design choices (e.g., clear labelling, secure defaults, trauma-informed flows). • Define necessary policy rules (e.g., Community Guidelines specific to the feature).	Detailed design specifications incorporating safety features, UI mock-ups demonstrating safety flows, draft policies, documented risk assessment.

#	Phase	SbD Focus	Activities	Output
3	Development/ Implementation	Ensuring safety features are built reliably and securely according to design specifications.	• Adhere to secure coding practices. • Ensure the robust and reliable construction of all designed safety features and mechanisms. • Implement necessary logging for moderation and incident review.	Functional code incorporating specified safety features and security measures.
4	Testing/ Quality Assurance	Verifying that safety features work as intended and testing for potential vulnerabilities or misuse scenarios ('abusability testing').	• Include comprehensive safety feature testing in Quality Assurance plans (e.g., Does blocking work fully? Is reporting intuitive? Do filters catch intended content?). • Conduct abusability testing – deliberately trying to break or misuse the product and its safety features. • Include comprehensive security testing (penetration testing, vulnerability scanning).	Test results confirming safety feature functionality and identifying any vulnerabilities found.
5	Launch/ Deployment	Ensuring a safe rollout, clear communication to users, and readiness for monitoring and response.	• Prepare clear user communications about new features, associated rules, and available safety tools. • Ensure moderation and support teams are trained and ready. • Implement phased rollouts where appropriate to monitor initial impact. • Final pre-launch safety sign-off.	Feature launched with supporting user documentation and prepared operational teams.

#	Phase	SbD Focus	Activities	Output
6	Maintenance/ Iteration	Ongoing monitoring, learning from real-world usage, responding to incidents, and iteratively improving safety measures.	• Continuously monitor user reports, feedback, and metrics related to the feature's safety. • Conduct regular Safety Incident Reviews when harms occur. • Utilize insights to update risk assessments, refine features, improve policies, and enhance moderation effectiveness. • Regularly review and update safety features based on evolving threats and user needs.	Updated risk assessments and testing, change impact assessments, improved safety features/policies, incident review findings, ongoing monitoring reports.



Key Questions for Each PDLC Stage

Learning from Failure for Better Prevention

When safety incidents occur, it presents an opportunity for improvement. Implementing structured Safety Incident Reviews is essential for capitalizing on this opportunity.

A thorough incident review process seeks to answer key questions:

- **What happened?** Clearly define the specific harm that occurred and identify who was impacted (users, specific groups, the product).
- **Why did controls fail?** Analyze which specific prevention or mitigation controls failed to function as intended, were bypassed, circumvented, or were simply missing altogether.
- **What were the root causes?** Look beyond immediate triggers to identify underlying contributing factors.
- **How could this have been prevented?** Assess how different design choices, stronger policies, or more robust processes could have potentially prevented this specific harm from occurring in the first place.
- **What corrective actions are needed?** Identify concrete steps to strengthen existing controls or implement new preventative measures based on the analysis.
- **What systemic changes are required?** Consider whether the incident reveals broader weaknesses in the PDLC, risk identification processes, training programs, or organizational culture that need addressing.

Activity 6: PDLC Safety Checkpoint Mapping

Setup & Role Assignment (5 mins):

- Divide participants into groups of approximately 6 people. (Facilitator Note: Aim for groups where each person can represent one PDLC stage).
- Each group member takes on the role of the "SbD Lead" for one stage of the PDLC: **1. Ideation, 2. Design, 3. Development, 4. Testing, 5. Launch/Deployment, 6. Maintenance/Iteration.**
- (Optional: Provide brief role cards outlining key responsibilities/focus areas for each stage based on Module 3 content).

Select Product Concept:

- Each group chooses **one** product concept to focus on for the activity.
- (Option A - Preferred if Activity 4 was expanded): Use the digital product concept and associated user personas the group developed during Activity 4 (Mapping Digital Personas).
- (Option B - Alternative): Choose from a pre-prepared list of concepts relevant to the local context (e.g., "A mobile app connecting local artisans directly with buyers," "A community platform for reporting infrastructure issues like water leaks or power outages," "An online tutoring service for high school students").

Activity 6 Continued: PDLC Safety Checkpoint Mapping

Phase 1: Individual Stage Brainstorming (15 mins):

- Working **individually,** each "SbD Lead" brainstorms 2-3 specific SbD activities, questions, safety checks, required inputs/outputs, or deliverables relevant o**nly to their assigned PDLC stage** for the chosen product concept.
- Encourage them to think about: What safety issues are most critical at this stage? What needs to be received from the previous stage? What needs to be handed off to the next stage? (Use sticky notes or sections on the group worksheet).

Phase 2: Group Integration Discussion (20 mins):

- Group members come together and present their stage-specific ideas sequentially (Ideation Lead first, then Design Lead, etc.).
- As a group, discuss:
  - How do the activities link together across the stages?
  - Are there gaps? (e.g., Did the Test Lead plan to verify a safety feature the Design Lead didn't specify?)
  - Are there overlaps or redundancies?
  - How can the plan be refined to ensure a smooth, continuous focus on safety throughout the entire lifecycle for this specific product?
- Arrange/Refine the points on the group's shared worksheet to show the integrated flow.

Phase 3: Plenary Share-Out (10-20 mins):

- Ask 2-3 groups to present their integrated SbD plan for their chosen product. They should briefly highlight key safety actions at each PDLC stage and explain how they connect

Facilitator's Note: Clearly explain the activity structure and the different roles. Provide or guide product concept selection. Manage time for both individual brainstorming and group integration phases. Circulate during group work to answer questions and prompt thinking about cross-stage dependencies. Facilitate the share-out session, drawing attention to effective integration strategies and the importance of each stage contributing to the overall safety outcome.

Learning from Failure for Better Prevention

The systematic risk identification and mitigation strategy outlined previously must be deliberately applied to address the specific risks of Technology-Facilitated Gender-Based Violence (TFGBV), gendered digital harms, online violence against women and girls. In the exercise above, where different team members took on SbD responsibilities in different stages of the PDLC, it is key that these leads bring in high awareness of this category of harms. Teams might also consider having a designated lead on ensuring SbD with gendered lens for the whole cycle.

Let's illustrate this application using the example of a user-to-user direct messaging (DM) feature:

1. **Identify TFGBV Harms**: What specific TFGBV risks could DMs facilitate? Examples include unsolicited sending of sexual messages or intimate images (cyberflashing, image-based abuse), persistent harassment or threats, stalking behaviors (e.g., using profile information found via DMs), sextortion, intimidation, etc.

2. **Identify Who/How Harmed:** Who is most vulnerable to these harms via DMs? Often women, girls, activists, journalists, or members of minority groups. How are they harmed? They experience fear, intimidation, silencing, psychological and mental health harms, reputational damage, threat of physical violence.

3. **Evaluate Impact & Prioritize:** Assess the likelihood and severity of each identified TFGBV risk. For instance, the non-consensual sharing of intimate images might be rated as having potentially High Severity, thus requiring high-priority controls. Persistent harassment might be High Likelihood but Medium Severity, still demanding significant attention. ***It is important to note – all risks need to be addressed, this exercise is meant to establish the order of priority and identify appropriate mitigation tools, not to ignore certain risks because they have low likelihood or have been evaluated to have low severity by the design team.***

4. **Develop Prevention/Mitigation Controls:** Brainstorm specific design, policy, and technical controls targeting the prioritized TFGBV risks in DMs:
   - Implement automated image scanning with sensitivity screens: Use technology (like AI/ML and potentially hashing against known CSEM/NCII databases) to detect potentially explicit or harmful images sent via DM, automatically blurring them and displaying a warning, giving the recipient control over whether they view the image. (Technical & Design Control)
   - Introduce a 'Message Request' system for non-contacts: Messages from users who are not mutually connected or followed land in a separate request folder with limited visibility (e.g., hiding full profile, disabling links) until the recipient explicitly accepts the request. (Design & Workflow Control)
   - Enhance in-chat reporting for abuse: Provide clear, easily accessible options directly within the DM interface for users to report specific messages or entire conversations for violations like harassment, threats, hate speech, or non-consensual imagery, ensuring these reports are handled appropriately according to policy and feedback is provided. (Design, Policy & Technical Control)

5. **Plan for Review & Iterate:** Continuously monitor DM usage for abusive patterns, analyze reports related to DMs, and refine controls based on effectiveness and user feedback.

A CASE STUDY: SbD and Compliance in the Kenyan Regulatory Landscape

While safety-by-design inherently pushes beyond minimal legal requirements, understanding the relevant **Kenyan regulatory landscape** provides a crucial baseline and reinforces certain SbD principles for stakeholders in that space. Adherence to Kenyan laws like the Computer Misuse and Cybercrimes Act (CMCA) and Data Protection Act (DPA) is mandatory and forms a critical part of responsible product operation. Compliance directly supports and operationalizes key SbD principles, including Provider Responsibility, User Empowerment, and Transparency.

Computer Misuse and Cybercrimes Act, 2018 (CMCA)

This Act directly criminalizes several actions often constituting TFGBV. Key relevant sections include:

- Section 27: Addresses **cyber harassment,** covering offensive, indecent, or threatening communication directed at individuals.
- Section 37: Targets the **wrongful distribution of obscene or intimate images**, crucial for combating image-based abuse.
- Other relevant provisions cover false publications (s. 22, 23), child pornography (s. 24), and identity theft/impersonation (s. 29) phishing s.30,

Implication for SbD: The CMCA provides legal backing for digital products and services to take action against specific harmful behaviors and offers avenues for legal redress, although practical enforcement can face challenges (5). Provider policies should align with, and ideally exceed, these legal prohibitions.

Data Protection Act (DPA), 2019

This Act is central for user privacy, trust and safety, aligning closely with several SbD principles. It establishes obligations for entities processing personal data of individuals within Kenya. Key aspects include:

- **Core Principles:** Mandates adherence to principles like Lawfulness, Fairness, Transparency (users informed about data use), Purpose Limitation, Data Minimisation (collecting only necessary data) Accuracy, Storage Limitation, Integrity & Confidentiality and Accountability.
- **User Rights:** Grants individuals rights such as accessing, correcting, and deleting their personal data, directly supporting the SbD principle of User Empowerment.

Implication for SbD: The DPA provides a strong framework for responsible data handling, which is key to user safety. SbD practices related to privacy controls, data security, and transparency directly support DPA compliance (21).

Activity 7: Day 1 Recap

Format: Individual written reflection, followed by optional brief sharing.

Instructions:

1. **Introduction**: Facilitator introduces this final activity for Day 1 as a personal reflection exercise to consolidate learning.
2. **Individual Reflection (10 minutes):** Ask participants to take a few minutes to quietly reflect on today's sessions and write down responses to the following prompts:
   - **3** Key concepts, ideas, or principles that were most impactful or new to you today.
   - **2** Potential ways you could start applying something learned today in your own work context or team discussions.
   - **1** Question you still have, or one topic you'd like to understand better as we move forward.
3. **Optional Sharing (3-5 minutes):** If time permits, invite a few volunteers to share one point from their reflection (e.g., their most impactful concept or a question they have).
4. **Collect Questions:** Encourage participants to leave their written questions (anonymously if preferred, e.g., on sticky notes) with the facilitator, who can review them and aim to address common themes on Day 2.

Facilitator Role:

Facilitator Role: Guide the reflection process clearly. Emphasize that this is primarily for individual processing but sharing is welcome. Manage time effectively. Collect any written questions.

MODULE

4

Deeper Dive Into
DESIGNING FOR SAFETY
Principles and Practices

MODULE 4

Deeper Dive into Designing for Safety - Principles and Practices

Learning Objectives

Upon completion of this module, participants should be able to:

1. Identify and categorize different types of technical and design interventions for SbD (Foundational, Enhancing, Prevention, etc.).
2. Describe specific SbD features and tools (e.g., Safe Defaults, User Controls, Verification Methods, Content Warnings, Blocking/Reporting, Privacy Controls).
3. Explain the function and importance of effective reporting mechanisms and trauma-informed design in reporting flows.
4. Analyze real-world case studies to evaluate their application of SbD principles and features.
5. Discuss how specific design choices can mitigate potential misuse (abusability).

Materials Needed

- Flipcharts
- Markers
- Handouts

Time Allotted

2 hours

Activity 8: Day 1 Recap

Request previously assigned participant (At the beginning of day 1, a participant should have been chosen to take notes for a recap session) to conduct a brief recap session.

Pair Up: Ask participants to turn to a person sitting near them to form pairs.

Pair Discussion (5-7 minutes): Instruct pairs to briefly discuss the following questions:

- "What was the single most important or surprising concept you learned about safety-by-design yesterday?"
- "Recall one of the 7 core SbD Principles. Why is that principle relevant in your work context?"
- "Briefly, what's a key difference between the user-focused Risk Assessment process and the attacker-focused Threat Modelling approach we discussed?"

Plenary Share-Out (5-8 minutes):

Facilitator invites volunteers to share one key insight or answer from their pair discussion with the larger group. Aim to get responses covering each of the three questions.

Translating Principles into Design

Based on the TSPA framework, and built on top of risk assessment and user feedback, the following technical and design interventions can be implemented. The following sections group various SbD interventions into broad categories to help illustrate the different ways safety can be embedded. While many features embody multiple principles or could fit into several categories, this grouping helps distinguish between different types of safety approaches:

- **Foundational Elements:** These are the basic building blocks that establish a baseline level of safety, set user expectations, and provide essential capabilities from the very start.
- **Safety Enhancing Features:** These are tools often added to improve safety in specific contexts, such as verifying user identity or securing communications.
- **Proactive Harm Prevention Technologies:** This category focuses on often technical or automated systems designed to detect, block, or mitigate harm before it reaches or significantly impacts users.
- **User Agency and Autonomy Tools:** These features specifically empower users by giving them direct control over their experience, interactions, content visibility, and personal data.
- **Content Moderation & Incident Reporting/Recourse:** These are crucial mechanisms for identifying and addressing harmful content or behavior that occurs in the digital tool or product, providing ways for users and the provider to respond to incidents, and offering avenues for appeal.
- **Transparency Features:** These tools aim to provide users with clarity on how the digital product or service operates, how decisions (like content removal) are made, and the overall safety performance.

Foundational Safety-by-Design Elements

Safe Default Settings: This ensures that there is always a baseline safety for all users on digital products and services, and provides a default level of protection before other interventions are made. This especially works for users who may not actively configure settings due to a myriad of reasons including low awareness of options, low digital literacy levels, lack of awareness about the configurability of a digital product or service, lack of time to change default settings, trust that settings are safe by default, etc. Examples include default higher privacy settings such as keeping profile details hidden from search engines or default private accounts for minors (22).

Opt-in/Opt-out Settings: Making privacy-protected choices central to a user's experience is fundamental. This can be achieved, and especially for higher risk features such as access to sensitive data from cameras or microphones or location data, by providing opt-in consent mechanisms for features with less security, with clear instructions or explanations. Users should not be left to opt-in to a safer experience; they should be given and option to opt-out of more private settings if they wish AFTER they have been given full information about the risks of lowering safety thresholds. This is the preferred method over setting up opt-out mechanisms as a default, which means features will be on until a user chooses to opt-out of them (23).

Clear Onboarding and Norm Setting: During the onboarding process, communication should be effective, concise, clear, and in language that users understand - focusing on community guidelines, terms of service and expectations for user behavior and conduct. This helps establish norms for safe interactions (23).

User Controls: The key to empowering users lies in providing them with granular, accessible, and intuitive controls over their experience. This includes tools like blocking or muting other users, managing privacy settings, controlling visibility of their own content and filtering content based on preference (24).

Transparent and Intuitive Design: Clarity in our design choices helps to build trust and empower users (23). User interfaces should be clear, easy to use and understand and avoid deceptive design elements that lead to the manipulation of users into making unintended choices otherwise called 'dark patterns' (23). Examples of deceptive elements include:

- Hard-to-Cancel Subscriptions (Roach Motel): Making it extremely easy to sign up for a service or mailing list but designing the cancellation or opt-out process to be deliberately confusing, lengthy, or hard to find.
- Pre-checked Consent Boxes: Automatically selecting options (e.g., agreeing to data sharing, signing up for marketing emails) using pre-checked boxes, relying on users overlooking them rather than making an active choice.
- Confirm Shaming: Using wording that guilts or shames the user into selecting the option the digital product or service prefers (e.g., offering choices like "Yes, I want discounts!" versus "No, I prefer paying full price").

Safety Enhancing Features

Identity Verification: These tools can be effective in fostering trust and accountability. They are often used in digital products and services that must understand who is using their products to ensure safe experiences for all users, such as ecommerce services and match making/ dating services (23). They ensure that all users, before being given access to spaces where interactions with other users can take place, pass through verification mechanisms to ensure that they are who they say they are while also protecting their data privacy.

Discussion:

Different types of services require different levels of user identity verification. Discuss what level of user identity verification is prudent, safe, and necessary in your company.

Age Verification and Assurance: These tools are used to estimate or verify a user's age, and are often specifically focused on protecting minors by restricting access to certain features or content and creating an environment where children can enjoy age-appropriate experiences (25). These methods vary in their effectiveness (e.g. just asking to confirm that a user is over 18 is not effective) and have implications for data privacy (when asking for birth date and full name), however they are a cornerstone of ensuring child safety online (26).

Discussion:

When and why is age verification important for your products? What could be an effective and safe measure for age verification?

Secure In-app Communication: These features ensure that a user's personal information, contacts, and other sensitive data is protected from accidental or deliberate abuse. Secured end to end encryption helps achieve this. Protecting users from sharing of personal information and contacts in order to connect with other people on a digital product or service is important. By providing secured (end to end encrypted) communication channels within an app can significantly reduce the risk of users sharing 'personal contact information' (23).

Discussion:

How might providing secure in-app communication change user behavior regarding sharing personal contact details? What are the potential trade-offs (e.g., usability, user preference) when deciding whether to implement this versus relying on users exchanging external contact info?

Content Warnings and Blurring: These tools protect users from being exposed to unsolicited and unwelcome content in their digital spaces, such as messaging apps. Developed in response to the rise of cyberflashing and utilized by companies such as Bumble, these tools obscure, blur, and place warnings on content that might be unwelcome, such as sexualized images. This allows users to be able to choose whether they would like to view such images, significantly reducing the risk of exposure (23).

Discussion:

How effective do you think content warnings/blurring are in mitigating harm versus outright blocking of certain content types? How might digital products or services define 'unwelcome' or 'sensitive' content consistently and fairly across diverse user bases?

Blocking and Muting: These tools and functionalities are effective for user empowerment, allowing individuals to own their experience and protect themselves from unwanted interactions. Effective design considerations include:

- **Comprehensiveness:** Ensure blocks are robust. Does blocking one account prevent interaction from other accounts potentially created by the same user? Does it prevent the blocked user from seeing the blocker's interactions with others?
- **Ease of Management:** Provide clear interfaces for users to manage their blocklists.

Addressing Pile-ons: Consider tools that facilitate mass blocking or temporary restriction of interactions during periods of intense, coordinated harassment (18).

Discussion:

Considering the design complexities mentioned (comprehensiveness, managing blocklists, pile-ons), what's the most challenging aspect of implementing truly effective blocking/muting? How can digital products and services balance robust blocking features with potential misuse (e.g., coordinated blocking campaigns used for silencing)?

Realtime Safety Alerts and Tools: These features, which include warnings relative to activity or interactions that a digital product or service may deem to be risky, including emergency assistance buttons and journey tracking with safety options for shared ride apps, enhance the users experience by introducing an element of immediate safety, which they themselves can act on and control (23).

Discussion:

For your specific products or services, what types of 'real time safety alerts' could be most beneficial or relevant for your users? How can these alerts be designed to be genuinely helpful and empowering without causing unnecessary user anxiety or 'alert fatigue'?

Tailored Safety-First Experiences: These tools create user experience that are customized or tailored and adaptable to the needs of individuals. These experiences can be adapted based on risk factors, such as displaying mental health support resources based on users searching for self-harm related content or adjusting content recommendations (23).

Discussion:

What ethical considerations arise when tailoring user experiences based on inferred 'risk factors' (e.g., potential inaccuracy, stigmatization)? What data might be needed to effectively tailor experiences (like showing mental health resources), and how can this data be gathered and used responsibly and transparently?

Parental Controls: These tools allow parents and guardians to be able to have granular controls over the experiences of those in their care. This can be achieved by allowing them to set content limits, feature restrictions, and active account monitoring (22). There are known limitations and shortcomings of parental controls, which could run the danger of shifting responsibility for safeguarding from the digital product or service to the caregiver, so these need to be built with safety-by-default principles as well.

Discussion:

How can parental controls be designed to genuinely empower caregivers without unfairly shifting the digital product or service's core safety responsibility? What are the key challenges in making parental controls effective yet respectful of older children's growing need for autonomy and privacy?

Proactive Harm Prevention Technologies

Automated Abuse Detection: Using emerging technologies such as AI and machine learning (text classifiers, image/video analytics tools), digital products or services can proactively detect and flag or remove content that violated safety policies such as hate speech, bullying, spam (24). The proactive nature of such tools allows digital products or services to mitigate harms even before users report it or get a chance to interact with it.

Hashing Technology: The creation of unique digital footprints, otherwise known as hashes of known illegal content, with a focus on the propagation of nonconsensual intimate images (NCII) and Child Sexual Exploitation Material (CSEM), terrorism related material and other sensitive or inappropriate content allows digital products or services the ability to detect and block re-uploads, in essence stemming the spread of such content (23).

Content Moderation Algorithms: These algorithms are used by social media platforms to detect and remove harmful content, such as hate speech, misinformation, and cyberbullying (27).

Digital Wellbeing Features: These features monitor and limit screen time, promote healthy usage habits, and provide alerts for excessive use. Many apps and tools, as well as service provides include these reminders and alerts to limit digital overwhelm, which can lead to poor mental health outcomes.

Cybersecurity Measures: Known tools such as advanced security protocols, firewalls, and anti-malware software protect users from cyber threats and data breaches.

User Agency and Autonomy Tools

Privacy Controls: A distinct category of tools that grants users fundamental agency over their personal information and online visibility. Best practices emphasize:

- **Clarity & Simplicity:** Use plain language, avoiding technical jargon, to explain what each setting does.
- **Granularity:** Offer users fine-grained options to control aspects like profile visibility, who can view their posts or stories, who can send connection requests or messages, and who can comment on their content.
- **Secure Defaults:** As per SbD principles, privacy settings should default to the most protective options (1).
- **Proactive Review:** Implement mechanisms like periodic 'Privacy Check-ups' to prompt users to review and confirm their current settings.

Tools for Content Moderation and Incident Reporting

Policy-Driven Moderation: Safety-by-design plays a key role in informing how policies can be developed in a clear and enforceable manner, based on a digital product or services content and behavior standards. While most digital products or services have human moderation as a primary tool, increasingly companies build moderation systems that balance automation and human review, essentially reducing the chances of harms being identified and dealt with before they affect the users (24).

Effective Reporting Mechanisms: In the case where, despite all efforts to pre-emptively deal with harmful content or interactions, incidences do occur, it is important to provide users with easy to find and use, intuitive tools for reporting issues they may encounter (24).

Appeals Processes: To enhance fairness and accountability, digital products and services should implement features that allow users whose accounts or content has been flagged, blocked or removed, to be able to appeal moderation decisions where they do not agree with these decisions (23).

Features and Tools to Increase Transparency and Accountability

In-Product Explanations: SbD mandates that as part of transparency, digital products and services should provide clear and easy to access explanations to users about how certain features work, such as recommendation algorithms or advertising systems (23).

Clear Rules and Enforcement Information: It is imperative that digital products and services make rules and the consequences for violation both accessible and easy to understand (23). Along with this, it is equally important to notify users about actions taken with regards to their content or accounts and on reports that they have raised using reporting tools.

Transparency Reports: regularly publishing aggregated data on content moderation actions, prevalence of harms and harmful content, user reports and appeals along with vulnerable group focused data is key when implementing effective SbD (1).

Fairness Measures: Techniques like fairness-aware machine learning and bias detection algorithms help ensure AI systems do not perpetuate or amplify biases in products and apps that rely on historical data to inform services for consumers (28).

Reporting

Effective user reporting mechanisms are important, acting as a channel for identifying harm and enabling mitigation. However, poorly designed reporting flows can be confusing, ineffective, or even re-traumatizing for users already experiencing distress (20). Best practices for designing safer reporting flows include:

- ***Accessibility***: Reporting options should be highly visible and easily accessible directly from the context where the harmful content or behavior is encountered (e.g., on a post, profile, or within a message) (10).
- ***Clarity and Specificity:*** Use clear, simple language and provide distinct categories that accurately reflect different types of harm (including specific forms of TFGBV). Allow users space to provide necessary context or details about the incident.
- ***Trauma-Informed Design:*** Employ non-blaming, supportive language throughout the flow. Provide clear 'quick exit' options. Allow users control over whether and how they receive follow-up notifications about their report to avoid unwanted reminders (18).

Efficiency for Addressing Mass Abuse: Design the system to handle situations involving coordinated harassment or 'pile-ons' efficiently, potentially allowing users to report multiple pieces of content or multiple accounts in a single flow (18).

Feedback and Appeals: Provide clear, timely feedback to the reporter regarding the status of their report (received, under review, action taken) and the outcome. Offer a clear and accessible process for appealing decisions if the user disagrees with the outcome (1). Consider features like user-facing report tracking dashboards (18).

Activity 9: Digital Safety Features Exploration – Perspectives (15 mins)

In small groups, think about a digital tool that you use often and select one as a group. Discuss its SbD features using questions below as a guide.

- What features does this product have that you have liked using?
- What features, now that you think about it, does the project have that you have been using without even noticing?
- What features discussed above are lacking in this tool or they exist but extra effort is required to activate them?
- What would make our user experience even better? Any features not on the list that you would create?

Then groups are given an additional task to think about:

Now imagine your … (different scenarios for different groups)

- 18-year-old niece/cousin
- Mother's sister who lives in a village outside Mombasa
- University friend who went on to be an anti-corruption reporter

… is using this technology. Place yourselves in their digital shoes - any differences or protections you would like to see?

Share what you have discussed briefly with the group. Which features or a combination of features are most important for you and for those in whose shoes you have spent some time?

Activity 10: Mini Case Study/Feature Redesign (15 mins):

- Introduce Scenario: "Imagine your product is adding a feature allowing users to create collaborative public playlists (e.g., for music or videos)."
- In small groups discuss:
  - What are potential safety risks, especially related to TFGBV? (e.g., Abusive playlist titles/descriptions, using collaboration invites for harassment, adding harmful content, brigading via playlists).
  - How could you redesign this feature or add controls to prevent/mitigate these risks? (e.g., Content filtering on titles, owner controls over collaborators/content, reporting for playlists/collaborators, privacy settings for playlists).
  - Facilitate brief share-out by asking groups to share one key risk and one mitigation/prevention idea.

Activity 11: Practicing Feature Redesign

Examining how safety-by-design principles are applied (or misapplied) in popular online services and products can provide valuable insights into their potential and limitations.

Platform-Level Examples

Major online digital products and services often publicly commit to SbD principles and implement various safety features, though the effectiveness and consistency of these efforts are subjects of ongoing debate and scrutiny.

- **YouTube:** YouTube has implemented a number of proactive safety features to help combat the spread of harmful information, CSEA, hate speech, misinformation and harassment (29). The platform uses machine learning to proactively detect harmful content and pairs that with human review. The platform also provides mechanisms to enforce age restrictions to ensure minors do not access content that might not be appropriate for them (YouTube kids) and employs 'quality principles' as a guide for children's content (29). Overall, the environment has been developed to give users granular control over the content that they watch. However, YouTube and other large providers have been the subject of ongoing criticism as a result of inconsistencies in enforcement, especially regarding hate speech and misinformation.

- **Meta (Facebook/ Instagram):** Meta introduced a message gatekeeping system to offer users protection from unwanted interactions on Instagram DM and Facebook Messenger, limiting the number of messages and the type of content a non-follower can send to a user. This restricts non-followers from sending any kind of media to users except for a text-based message, which is subject to approval from the recipients before viewing (9). They have also implemented tools to support youth safety by making sure that accounts for people below 16 are private by default and offer parental controls (22). However, Meta has been criticized for failing to effectively act on reported hate speech and abuse happening via direct messages (14). They also recently shifted from proactive content moderation towards user-driven annotation systems – straying from the core tenets of SbD (30).

- **TikTok:** TikTok made efforts to reduce discoverability of search results that violate community guidelines (9). In a public report, TikTok noted that they had proactively removed 92% of hate speech and hateful content before anyone reported it, with 87% of it being removed within the first 24 hours of upload (9). However, like their counterparts, they have been subject to criticism with regards to enforcement failures around harmful content (14).

- **X (formerly Twitter):** While this platform has been notorious for inaction towards reported abuse (14), they have made shifts towards 'Community notes' which represents a decentralized approach to content moderation and adding context. This differs significantly from traditional, centralized content moderation models where employees typically make direct decisions about content removal or labelling. Concerns about the quality of such moderation models persist. (31).

- **Bumble (and other dating apps):** Bumble developed the 'Private Detector', an AI driven tool that blurs potentially harmful or inappropriate images shared in chats, giving users control over what they choose to see (9). Dating products are predisposed to the potential for high interpersonal harm and have implemented strong verification steps, and robust reporting and blocking tools with the goal of reducing the chances for harm to take place (23).

- **Gaming products (e.g. Twitch and Overwatch):** Gaming products such as Twitch face immense challenges around toxic behavior, harassment, and misogyny (9). In response, products such as Overwatch have incorporated targeted reporting mechanisms, with reporting options that are particularly instructive in that they describe what constitutes a violation of community guidelines and what does not (32).

- **Children's products and services (e.g. Roblox):** Platforms used by children are under constant pressure to implement strong SbD measures to curb harms before they occur. However, research has shown that some design choices that optimize engagement have the potential to lead to unintended consequences such as exposure to harmful content, inappropriate contact with adults, and exposure to risks even when age restriction has been implemented (33). Roblox has Community Standards that outline what behavior is and isn't acceptable on the site. They also have moderators who review and remove inappropriate content. They also have a range of parental controls and safety settings that can help keep your child safer on the site (34).

Feature/Sector Specific Examples

Going beyond product-wide implementations of SbD, specific features have demonstrated SbD principles in action:

- **Payment Description Moderations:** In Australia, several banks have taken proactive measures to detect and block abusive messages sent via the transaction description fields in digital payment systems (35). This demonstrates how SbD principles can be applied in digital products and services that are non-social in nature to address financial abuse, a well-known form of domestic violence (35).
- **Ride-Sharing Safety Tools:** Platforms such as Lyft and Uber have taken steps to ensure both passenger and driver safety. These interventions include the addition of in app emergency buttons, real-time location sharing with contacts in a trusted list, and dedicated and easily accessible support channels (23).
- **AI for Harassment Detection:** Some social media platforms, such as Instagram and YouTube have taken steps beyond just using AI to remove harmful content to proactively identifying patterns associated with harassment or abusive language, with some using user prompts or limiting content visibility (9).
- **Default Private Accounts for Minors:** Platforms such as Instagram and Facebook have made moves towards making accounts of users below the age of 16 private by default with parental control features, in essence reducing the chances of unwanted contact or exposure using SbD principles (22).
- **Non-Consensual Intimate Image (NCII) Prevention:** The existence of tools such as StopNCII.org, which allow users to proactively hash their intimate images so that other digital products and services can block them from being shared without consent, highlight how technology driven SbD can make a global difference. Ofcom recognizes this as good practice (13). Platforms actively using hashing technology (creating a unique digital fingerprint - a "hash" for an image to allow for detection by 'forward' blocking tools) are able to stem the spread of CSEM and other harmful content (23).
- **Removing Unwanted DMs:** Communications apps and services, such as Instagram and Bumble, are implementing design features to limit unsolicited or potentially harmful DMs, particularly for vulnerable users, such as filtering message requests from unknown senders or blurring potentially explicit images (9).

Reflection Question:

Which of these are doable in your tool/product? Do you have concerns or thoughts about these approaches?

MODULE

5

BUILDING SbD CULTURE

MODULE 5

Building SbD Culture

Learning Objectives

Upon completion of this module, participants should be able to:

1. Define 'safety culture' within the context of technology organizations and explain its critical importance for sustainable SbD implementation.
2. Identify and discuss common tensions between safety objectives and other organizational priorities like usability, innovation speed, freedom of expression, and privacy.
3. Describe key practical strategies for actively building and strengthening a safety-first culture across an organization.
4. Recognize common operational barriers that hinder SbD adoption and culture change.
5. Outline a structured approach with actionable steps for initiating or improving SbD adoption within their own teams or organizations.

Materials Needed

- Flipcharts
- Markers
- Handouts

Time Allotted

2 hours

What is Safety First Culture?

Within the context of tech products, and services, safety culture represents shared values, beliefs, attitudes, and consistent behaviors concerning user safety that exist at every level of an organization. It brings about an operational environment where the safety, dignity, and overall well-being of users are not just considerations to check off the list but are actively and consistently prioritized. This prioritization must extend across all relevant functions from product management and engineering to design, marketing, legal, customer support, policy enforcement, and trust and safety operations (9).

Why Does Culture Matter?

Without an ingrained organizational culture, any effort to create SbD practices and protocols can deteriorate quickly. Culture ensures that these practices and protocols are sustained, that commitments are not just stated but are fulfilled, and that all staff—from leadership to designers—understand, value, and uphold their roles in the process. In summary:

- **Consistency and Reliability:** Safety considerations are applied uniformly and reliably, even when faced with competing priorities or time pressures.
- **Proactivity and Prevention:** Teams and individuals think ahead, anticipate potential risks, and proactively build safeguards into products and features from the earliest stages.
- **Holistic and Effective Solutions:** Collaboration and communication enable creation of more comprehensive, robust, and well-rounded safety solutions that benefit from different perspectives and expertise.
- **Psychological Safety and Continuous Improvement:** Employees at all levels feel secure enough to voice concerns, report near-misses or potential vulnerabilities, and suggest improvements without fear of blame or negative repercussions.
- **Sustainability and Resilience:** SbD is embedded in organizational values, processes, and expectations, making implementation resilient to changes caused by staff turnover, shifting market pressures, or changes in short-term business priorities.

Challenges and Tensions Between Safety and Other Design Priorities

Successfully implementing SbD involves acknowledging and addressing tensions arising from competing priorities and values that have implications for product design. Some of these tensions are perceived and others are rooted in design constraints. They often require shifts in awareness and attitudes within the organization.

Brainstorm and Discussion: Which tensions can participants name?

Tension with Usability and User Experience

Introducing safety measures can sometimes be perceived to conflict with the UI/UX goals of seamlessness and ease of use. For example, strict age verification processes which are important for child safety can add steps to the onboarding process and may be perceived as being intrusive or inconvenient for users (26). Adding features such as warnings or delayed posting or sharing of content to curb impulsive harmful behavior can be perceived as an interruption in user flow (36). Tools like content filters can restrict the upload of legitimate content as well.

The solution to this challenge is to design features that are effective yet intuitive enough to empower users without making product use cumbersome (23). Achieving this requires heightened awareness of ‘dark patterns’, such as hidden costs/ drip pricing, roach motel/forced continuity and others, that exploit usability to manipulate users into unsafe choices (23).

Tension with Innovation and Creativity

Tech culture around innovation requires that innovators 'move fast and break things' in order to be successful in the market, which is a direct contrast to the risk-averse nature of SbD which requires digital products and services to move thoughtfully (1). In-depth risk assessments, red teaming, and user testing means potentially delayed deployments to address safety issues. These additional steps have the potential to increase development costs, slow down innovation, and hinder a product's ability to compete, more so for startups.

The solution to this challenge is to communicate the benefits of the safer product as a competitive advantage to customers and shareholders. This will bring about benefits like user trust and retention, brand reputation enhancements, and reduced costs associated with incident response. (23).

Tension with Freedom of Expression

Perhaps the most contentious of all of the trade-offs is striking a balance between the protection of freedom of speech and safety (30). Effective implementation of SbD principles for digital products and services means implementing features such as content moderation, or making deliberate design choices like algorithmic downranking that might limit the visibility of certain posts or content and might be perceived as censorship or restriction of legitimate yet controversial content (37). Creating a clear distinction between where freedom of speech ends and harms begins has been inherently difficult, with the interpretation of this 'line' varying from product to product (38).

Features, such as content filtering, that allow users to see what they choose offers one way to navigate these complexities, instead of imposing top-down restrictions (39). However, even with such interventions, there is still the risk that SbD could be misused to suppress dissent or enforce viewpoints that further specific authoritarian or corporate interests (37).

The long-term solution to this challenge is digital citizenship and digital rights education and creating a culture of positive digital engagement. More immediate steps in this direction include partnerships with civil society and the utilization of tools (such as lexicons of abusive terms cocreated with survivors of abuse) to inform content monitoring and moderation.

Tension with User Privacy

Efforts to enhance safety can sometimes conflict with user privacy. Certain safety measures necessitate increased data collection, monitoring, or processing. For instance, proactively scanning user communications or content for CSEA or grooming behavior involves analyzing private data (40). As an example, in order to detect patterns associated with harassment, behavior analysis might be required, while implementing age assurance methods might necessitate the collection of sensitive personal information such as government identification documents or biometric data which raise concerns around privacy (26). Such instances create a tension between the fundamental right to privacy and digital safety measures.

The strategy to overcoming these challenges includes emphasizing the equal weight of users' fundamental right to safety, prioritizing privacy-preserving safety techniques, and adhering to strict data protection principles like minimal use and purpose limitations while complying with relevant regulations.

Reflection Question: Are there other design tensions that we have not discussed, or any reflections on these that you would like to share?

How To: Culture Building Strategies for SbD

Building a genuine safety culture requires ongoing effort and commitment across the organization:

Visible Leadership Commitment: Organizational leadership must champion safety as a core value and strategic priority. To do so they must communicate its importance to all teams, allocate necessary resources and establish clear governance structures that hold leaders and teams accountable for safety outcomes (1).

Employee Involvement and Empowerment: There must be a clear expectation of safety being the responsibility of all staff in an organization. This is complemented by providing mechanisms for reporting safety concerns and establishing roles like safety leads within different teams.

Cross-Functional Collaboration and Communication: As an organizational commitment, SbD frameworking in the organization might involve breaking down established silos and implementing processes that require collaboration between teams.

Continuous Training and Awareness Building: Achieved by providing regular and role-specific training on SbD principles, TFGBV awareness, and other safety related trainings to all relevant staff (9) and sharing learnings from past safety incidents.

Integration into Standard Policies, Processes, and Workflows: Embedding safety checkpoints, considerations, and requirements into existing organizational processes is key (PDLC stage gates, design review templates, Quality Assurance testing protocols, feature requirements documentation, project kick-off meetings and others) (41).

Alignment of Incentives and Recognition: Ensuring that performance metrics, reward systems, and promotion criteria do not solely prioritize speed, growth, or engagement metrics at the expense of user safety. Actively recognize and reward teams or individuals who demonstrate leadership or make significant contributions to improving user safety.

Fostering Representation of Experiences and Perspectives: Actively ensuring that different perspectives and backgrounds are represented and considered by technical teams, product leadership, and decision-making bodies is important. A broader range of perspectives is crucial for identifying a broader range of potential risks and developing more responsive and effective safety solutions (20).

Activity 10: Culture Step Brainstorm

Individual Reflection (2-3 minutes):

Ask participants to individually think about the following question:

- "Considering the strategies we've discussed and your own work environment, what is one small, specific, and actionable step your team or organization could take in the next month to begin strengthening its safety culture?"

Write it Down:

Instruct participants to write down their single step on a sticky note or piece of paper. Encourage specificity (e.g., instead of "More training," maybe "Propose a 15-minute discussion on TFGBV risks in our next product meeting").

Pair/Trio Share (3-5 minutes):

Ask participants to briefly share their actionable step with one or two people sitting nearby.

Plenary Share-Out (5-7 minutes):

Facilitator invites several volunteers (or conducts a quick round-robin if time permits) to share the specific step they identified with the larger group.

From Theory to Action: Practical Steps for SbD Adoption

Implementing safety-by-design (SbD) principles and creating the necessary cultural shifts often encounters practical challenges within organizations that go beyond addressing the tensions described above. This section provides a pragmatic roadmap with actionable steps to embed SbD effectively.

Practical Steps for SbD Adoption

Recognize Operational Barriers to SbD Adoption

Organizations embarking on the SbD journey may face several common obstacles:

- **Resistance to Change:** Attachment to established workflows, or the perception that prioritizing safety might impede development speed or stifle innovation, particularly clashing with prevalent "move fast" cultures.
- **Lack of Awareness and Understanding:** A potential gap in clear understanding across various teams regarding core SbD principles, the specific nature and impact of digital harms like TFGBV, or the individual roles employees play in promoting a safer environment.
- **Conflicting Priorities and Metrics:** Intense organizational pressure to prioritize short-term growth metrics, user engagement figures, or feature release velocity over potentially longer-term, foundational investments in safety infrastructure and preventative measures.
- **Absence of Measurements of Safety Outcomes:** Defining and measuring 'Safety' and 'Harm' can be challenging. While it may seem clear to some, harm is often context-dependent and what one group of users finds to be harmful might not be perceived so by another (8). At the same time, different types of harm may vary in severity and prevalence (23). This makes it difficult to set universal goals.
- **Organizational Silos:** Ineffective communication channels or a lack of established collaboration mechanisms between key departments (e.g., Product Management, Engineering, Policy, Trust & Safety, Legal), which can hinder the holistic integration of safety considerations.
- **Unclear Ownership and Accountability:** Ambiguity surrounding who holds the ultimate responsibility for driving, overseeing, and being accountable for the implementation and outcomes of SbD initiatives.
- **Perceived Regulatory Gaps and Uncertainty:** Although SbD aims to exceed compliance, the absence of clear, stringent external mandates can sometimes make it more challenging internally to justify investments in proactive, preventative safety measures that go beyond legal minimums (5).
- **Resource Constraints:** Finally, it is important to recognize that implementing SbD at scale requires input and investment of time and expertise (32). This training and support should be provided in follow up and should attempt to reduce these constraints; understanding that this deliberate focus on SbD could be hard for smaller organizations, startups, and civic tech.

In an environment where regulatory space around safety seems to still be primarily focused on voluntary compliance before moving to more stringent regulatory requirements, stakeholders should make sure that smaller players have access to knowledge, skill-building, and tools to compete with large tech companies (1).

Utilize a Structured, Iterative, and Integrated Approach.

The following steps provide a practical framework:

1. **Secure Leadership Buy-in and Sponsorship:** There is a compelling business and ethical case for SbD, which can be specifically tailored to the organization's priorities, values, and mission. Crafting that case could be the first step – it would utilize data, including risk assessments, incident reports, user feedback, and relevant statistics on harms like TFGBV to clearly demonstrate the need for SbD and its potential benefits. The goal is to demonstrate active, visible support and sponsorship from senior leadership to secure the necessary resources and organizational visibility for SbD initiatives.

2. **Start Small and Demonstrate Value (Pilot Approach):** Avoid attempting a complete organizational overhaul immediately. Instead, select a specific, manageable pilot project/tool/app/product. This could also look like integrating SbD into a single new feature, redesigning a known high-risk area of an existing product, or addressing one key safety risk identified during initial assessments. The goal is to apply SbD principles within this defined scope, measure the impact, both qualitatively (user feedback, team experience) and quantitatively (relevant safety metrics), and use the results to discuss the value of SbD, build internal credibility, and lay the foundation for broader implementation.

3. **Integrate, Don't Isolate:** Weave safety considerations directly into existing development, product management, and operational policies, protocols, and processes rather than creating entirely separate and potentially burdensome safety checkpoints. Modify existing templates, checklists, and review stages to include safety prompts and checks.

4. **Provide Training, Resources, and Tools:** Equip teams across the organization with the necessary knowledge, skills, and resources. Share learnings from foundational training sessions. Advocate for and provide ongoing, role-specific education on topics such as understanding harms, risk mitigation, development practices, privacy engineering, internal safety policies, and so on.

5. **Assign Clear Responsibility and Accountability:** Ensure clear ownership for driving and overseeing SbD initiatives. This may involve nominating specific individuals or establishing a dedicated cross-functional team or steering committee with a clear mandate. Alternatively, this could involve revising current job descriptions to add safety-related responsibilities.

6. **Build Collaboration and Continuous Learning:** Implement mechanisms and forums that encourage communication, knowledge sharing, and problem-solving between all relevant teams.

7. **Measure and Communicate Impact:** Define, track, and regularly report on relevant safety metrics. These might include reductions in specific types of harmful content, trends in user reporting rates and satisfaction levels, measured effectiveness of specific safety features, or changes in user perceptions of digital product and service safety.

Note: Many digital products and services use proxy metrics such as the number of pieces of content detected that are in violation of policies to measure advances in safety, but they report struggling to establish direct causal links between SbD features and harm reduction (23). Lack of standardization of reporting methods and tools across industries makes comparison and benchmarking problematic from a big picture perspective (40). Regardless, measuring SbD efficacy is possible. We will work towards discussing how to set metrics in our practical sessions. Remember – what we measure – we improve!

MODULE

6

EMERGING TECHNOLOGIES

& Workshop Bridge

MODULE 6

Emerging Technologies & Workshop Bridge

Learning Objectives

Upon completion of this module, participants should be able to:

1. Identify novel safety risks associated with emerging technologies (AI, VR/AR, IoT)
2. Discuss potential applications of SbD principles to mitigate risks in these emerging technologies.
3. Recognize the need for dynamic and adaptive SbD frameworks.
4. Consolidate learning through a tabletop simulation applying SbD concepts.
5. Understand the connection between the foundational training and the subsequent practical prototyping workshops.
6. Recall the goals and focus of the upcoming prototyping workshops.

Materials Needed

- Flipcharts
- Markers
- Handouts

Time Allotted

1.5 hours

From Theory to Action: Practical Steps for SbD Adoption

The evolution of technology is rapid. Technologies such as artificial intelligence, virtual reality and the Internet of Things present opportunities to expand human capabilities and the tech landscape but also present new avenues for harm and risk, including TFGBV. The proactive nature of SbD also requires that its principles and practices remain adaptable to the unique challenges and risks that present themselves with new technologies. This module reviews examples of known risks in these technologies and explores potential avenues for SbD integration.

Artificial Intelligence and Generative AI

Novel Risks: While they have the potential to bring immense benefits to human productivity, they also have the potential to amplify existing harms. For example, in AI models built for moderation, recommendations, or access to opportunities, algorithmic bias (whether intended or not) can lead to discriminatory outcomes (42). Additionally, AI enables faster and more sophisticated creation and spread of manipulative information, such as deepfakes, NCII, and CSEM on a massive scale (31). It also provides tools for abusive tactics such as cyber-mobbing and tailored hate speech in large volumes. The general 'black box' nature of AI technologies and a lack of transparency also makes auditing such systems for safety extremely difficult (43).

SbD Application: For the effective application of SbD in tools and products that are powered by AI or utilize AI elements, like chat-bots, service providers must remember to embed safety, human rights principles, and ethical principles into the entire lifecycle of the development. This means including these considerations from the data collection, cleansing and feature selection stages, into the model training and validation stages and to the deployment and maintenance/iterative stages (42). This involves conducting comprehensive risk assessments on potential AI-driven harms, developing methods to detect and correct bias, and building transparency into the inner workings of AI models (30).

These solutions can be approached in many ways and across the board—including by working with stakeholders and industry players through pledge initiatives to uphold child safety principles, and by adhering to safety standards focused on women and girls and other vulnerable groups. AI can, if designed carefully to avoid causing further unintended harms, be used to create sophisticated SbD tools like content detection systems and to power content moderation tools (24).

Virtual Reality (VR), Augmented Reality (AR), and the Metaverse

Virtual Reality (VR) and Augmented Reality (AR), which are often associated with the idea of the metaverse (the metaverse is a concept for a persistent, shared, 3D virtual space or network of virtual worlds where users, often represented by avatars, can interact with each other and digital environments in real-time for socializing, work, commerce, and entertainment), create entirely different and immersive types of digital experiences, also with their own unique opportunities and risks.

Defining AR, VR, and XR

For clarity, extended reality (XR) is an encompassing term that covers any sort of technology that is used in altering reality by adding virtual/digital elements to the physical or real-world environment to any extent, leading to a blurring of lines between the physical and digital world. XR includes AR, MR, VR, and any technology, present or future, that exists or might exist within the virtual continuum (i.e., the full spectrum of possibilities between the entirely physical world and the fully digital world or virtual environment) (44).

Augmented reality (AR), on the other hand, can be defined as any technology that superimposes digital elements into a real-world environment. Virtual reality (VR) refers to any technology that allows for the creation of a fully immersive digital environment. In this experience, the physical or real-world environment is completely blocked out (44). Lastly, mixed reality (MR) refers to any technology that allows not only for the superposition of digital elements into the real-world environment, but also their interactions. With MR, users can see and interact with both the digital elements and the physical elements.

Distinguishing AR, MR and VR (44)

Novel Risks: AR and VR have created an environment for deeply immersive experiences, which can make harms such as harassment, assault, or exposure to disturbing content more visceral and traumatic (14). These technologies also generate an immense amount of data, ranging from biometric data, psychological responses, and environmental data, that poses significant privacy risks that are not addressed by existing frameworks (45). Due to their real-time and dynamic nature, traditional mitigation strategies such as content moderation become extremely challenging to execute effectively (45). Concerns also exist around the potential for these environments to facilitate forms of CSEM, grooming, and extremist recruitment (14). Additionally, access to these virtual environments can be challenging for persons with disabilities (42).

SbD Applications To effectively apply SbD to VR and AR, it is important to take a holistic approach, from hardware such as headsets (unlike audio headphones or AR glasses that overlay digital information onto your surroundings, VR headsets fully replace your real-world view with a computer-generated, immersive 3D environment using stereoscopic displays and head-tracking technology) and controllers to software products, application designs, and governance rules of virtual spaces (42). This includes building robust user controls for managing interactions (blocking, muting, personal space bubbles), developing new moderation techniques suitable for immersive environments, implementing privacy-enhancing technologies, and strong data governance frameworks (data minimization, purpose limitation, user control over biometric data) (45). Ensuring accessibility features are integrated and establishing clear community standards and enforcement mechanisms, as well as having intuitive and well labelled report, pause, and leave buttons is critical. SbD principles should also inform the design of avatars and interaction mechanics to prevent harmful behavior. Additionally, VR and AR offer potential for safety applications, such as immersive safety training simulations for hazardous environments (46).

Going beyond individual tools, we encourage companies to collaborate on human-centric interoperability standards that prioritize user safety and rights across the virtual reality ecosystem (45).

Internet of Things (IoT)

We live in an increasingly connected world, with devices in our homes and communities increasingly able to connect to the internet and to each other. The Internet of Things (IoT) refers to the network of everyday physical objects—from home appliances and wearables to industrial equipment and city sensors—that are embedded with software, sensors, and connectivity, enabling them to collect and exchange data over the internet. This brings about a few distinct safety and security challenges.

Novel Risks: Security vulnerabilities in IoT devices are frequently exploited, potentially leading to data breaches or allowing devices to be controlled remotely for malicious purposes (37). A particularly concerning trend is the weaponization of everyday connected devices (smart speakers, cameras, locks, thermostats) by perpetrators of domestic and family violence to monitor, harass, control, and abuse their victims (9). The constant data collection by IoT devices also raises significant privacy concerns (47). In industrial Internet of Things (IIoT) settings, interconnected systems and automation introduce new risks to worker safety if not managed properly (48).

SbD Applications: SbD for IoT necessitates embedding robust security and privacy protections into devices from the design stage (49). This includes secure default configurations, user options for customized security measures like app “hiding”, mechanisms for secure software updates, strong authentication,

and encryption of sensitive data. Crucially, designers must anticipate potential misuse scenarios, such as domestic abuse, conduct rigorous threat modelling, and build in features that mitigate these risks—for example, providing clear indicators when a device is recording, allowing users easily to revoke access permissions, or designing interfaces that make shared control transparent and manageable (49). In industrial contexts, SbD involves designing collaborative human-machine systems with safety at the forefront, using sensors and AI for hazard detection and prevention, and ensuring robust safety protocols for automated processes (48).

Anticipating Novel Harms and Evolving SbD

To be able to anticipate these dynamic, complex, and emerging risks means ensuring that SbD frameworks do not remain static. They must be just as dynamic as risks, as well as adaptable and forward-looking to address unforeseen harms as new technologies emerge. To do this requires:

- Continuous research, learning, and development
- Multistakeholder dialogue and industrial working groups
- Adaptive frameworks for integration of new opportunities and mitigation of new threats

Designing With Users: Introduction to Co-Design

Before we do some workshopping to bring this all together, there is one more important concept to cover. Co-designing with users for their safety. Co-design is a participatory approach that fundamentally redefines the relationship between designers/developers and users. It treats users and community members not just as subjects of research or recipients of finished products, but as active partners and experts in their own lived experience. In a co-design process, these individuals have an opportunity to collaborate with designers, researchers, product managers, and engineers throughout various stages of the design and development lifecycle (50).

The value proposition of employing co-design within the context of SbD is significant:

- **Accessing Lived Experience Insights:** Users, particularly individuals who have directly experienced digital harm or belong to groups disproportionately targeted by abuse (such as women, journalists, activists, or minority groups), possess invaluable, nuanced insights. They understand the realities of risks, the tactics used by malicious actors, the vulnerabilities within digital products and services, and what truly contributes to a feeling of safety online (18).
- **Ensuring Relevance and Usability:** Co-design helps guarantee that safety features, policies, and interventions are not just theoretically sound but actually address real user needs and pain points. It ensures solutions are designed in ways that are smart, accessible, and practical for different groups of users to use effectively in complex real-world situations.
- **Building Trust and Ownership:** Engaging users respectfully and collaboratively in the design process can cultivate trust between the platform/product/tool and the customer base. It can increase user buy-in for safety initiatives and potentially lead to greater awareness, adoption, and effective utilization of available safety features.
- **Identifying Unintended Consequences:** Users participating in co-design can often identify potential negative side effects, unforeseen risks, or possibilities for misuse associated with proposed safety features that internal design teams might not be able to see.

Ethical Engagement Principles: Doing Co-Design Responsibly

Ethical considerations are not optional add-ons, they are important and must guide every decision and interaction throughout the engagement process. The following core principles MUST be adhered to (51):

1. **Informed Consent**: Provide potential co-design participants with clear, simple, accessible, and comprehensive information before they agree to participate. This must cover the purpose, procedures involved, potential risks and benefits, expected time commitment, how their data will be used and protected, and who to contact with questions.
2. **Do No Harm (Prioritize Safety):** The physical, psychological, emotional, and digital safety and well-being of participants must always be the top priority, taking precedence over design objectives. Actively anticipate potential risks of distress or re-traumatization, particularly when discussing potentially upsetting experiences and harms.

3. **Confidentiality and Privacy:** Explain clearly and transparently how participant data (including personal information, responses, recordings) will be collected, stored securely, who will have access to it, and how it will be used. Implement data security measures to prevent unauthorized access or breaches. Anonymize data wherever possible and appropriate, ensuring this is clearly communicated and agreed upon with participants.

4. **Respect, Dignity, and Sensitivity**: Treat all participants as experts in their own experience, genuinely valuing their knowledge, perspectives, and contributions to the design process.

5. **Data Minimization:** Adhere strictly to the principle of collecting only the data that is absolutely necessary for achieving the specific, stated purpose of the engagement.

Trauma-Informed Design

Finally, SbD is operationalized to address real harms. Some of these harms are severe—online harassment can include death and rape threats, videos and images containing graphic violence, and hate speech. Examples of harms that occur as a result of doxxing include physical violence. This is traumatic content to discuss and analyze, and it is re-traumatizing for those who have experienced abuse in the past.

Therefore, the design of user interactions, particularly those related to safety, requires sensitivity. Trauma-Informed Design provides a framework for this. It highlights that users may have diverse experiences with trauma (including past or ongoing experiences of violence, abuse, or discrimination like TFGBV). It aims to design products, features, interfaces, and communication styles in a way that actively avoids causing further harm or re-traumatization while simultaneously promoting user agency, healing, and a sense of safety (20); (50).

Trauma-Informed Principles in Practice

Implementing trauma-informed design involves adhering to several core principles and recognizing that there is no one-size-fits-all solution (50,52):

- **Safety (Physical and Psychological):** Create an environment where users feel secure. Examples: clear 'Quick Exit' buttons on sensitive reporting pages, predictable and consistent UI flows, transparent communication about data security measures, avoiding victim-blaming language.
- **Trustworthiness and Transparency:** Build user confidence through reliability and clarity. Examples: being honest and clear about data use policies and moderation processes; following through reliably on actions promised (e.g., providing feedback on reports); using simple, unambiguous language.
- **Choice and Control:** Maximize user agency and autonomy. Examples: allowing users to control the level and type of notifications they receive (e.g., about report status), providing clear options for anonymity where safe and appropriate, making privacy settings easy to understand and manage, ensuring consent processes are truly informed and voluntary.
- **Collaboration and Mutuality:** Value user input and shared decision-making. Examples: meaningfully involving users (especially those with lived experience of harm, including survivors) in co-design processes (as discussed above), gathering feedback in respectful and accessible ways, acknowledging power dynamics.
- **Empowerment and Skill-Building:** Support user resilience and capacity. Examples: providing easy access to relevant support resources, helplines, and safety information; offering clear instructions and educational materials; ensuring features are accessible to users with diverse abilities and digital literacy levels.
- **Cultural Humility and Responsiveness:** Recognize and respect diverse backgrounds, identities, and experiences. Examples: avoiding jargon and culturally-specific idioms; offering multilingual support where feasible; being mindful of diverse cultural contexts related to trauma, disclosure, and help-seeking.
- **Peer Support:** Recognize the value of connection with others with similar experiences (where appropriate and safely facilitated). Example: designing a feature where users who have experienced similar types of online harm (e.g., specific forms of harassment or TFGBV) can opt-in to join private, well-moderated discussion groups within the digital product or service to share experiences, coping strategies, and offer mutual support, with clear guidelines and safety protocols.

- **Strengths-Based Approach:** Focusing on resilience and coping mechanisms alongside risks. Example: creating sections within the digital product or service that not only detail risks but also prominently feature guides on digital literacy, online self-care, developing digital resilience, and stories (with consent) of how others have successfully navigated online challenges.
- **Minimal Data Collection:** As a core privacy principle, only collecting data that is strictly necessary reduces potential harm if data is breached or misused (50). Example: when a user is reporting harm, clearly distinguishing between mandatory fields essential for investigation (e.g., link to offending content) and optional fields (e.g., detailed emotional impact, unless the user wishes to share for support linkage), explaining why each piece of information is requested.

Discussion:

As you have already noticed, there is a lot of overlap between SbD principles, trauma-informed design, and ethical co-design principles. Please turn to your neighbour and, in pairs, discuss why this is the case. Feel free to share your insights with the rest of the group.

Simulation & Linking to Prototyping Workshops

To consolidate the concepts learned across the modules and bridge theory with practice, we will now move on to carry out a hands-on simulation designed to apply integrated SbD thinking. This workshop will also lay the foundation for the practical prototyping workshops that will follow this training.

Activity 11: SbD Tabletop Simulation

Setup: Participants are divided into small working groups (5 to 6 people—can use previously assigned groups), equipped with materials like flip chart paper, whiteboards, or digital collaboration tools.

Scenario Example: A clear, concise scenario is presented to the groups. For instance:

"Your group is a digital platform company planning to launch a new 'Community Forum' feature integrated into your existing platform. This feature will allow registered users to create discussion topics publicly, post questions or comments within those topics, and reply to each other's posts."

Instructions and Key Questions:

Groups are given a specific timeframe (e.g., 20-30 minutes) for discussion and tasked with addressing key questions that draw upon concepts from all preceding modules:

1. **Outline 3 to 4 of the most basic non-safety related steps your company would take to launch any product.**
   - You can use actual processes from your real work or think of processes you might use if your group's company and scenario were real. Examples include brainstorming a new product, developing a beta version of the product, or collecting feedback from a test audience.

2. **Identify Key Risks:**
   - What are 2-3 primary potential TFGBV or other significant safety risks specifically associated with this public forum feature (e.g., targeted harassment within discussion threads, organized spreading of hate speech or misinformation, doxing users based on their forum activity, coordinated brigading/trolling campaigns)?
   - Who might be particularly vulnerable to these risks in the context of a public forum?

3. **Prioritize SbD Principles:**
   - Which 1-2 core SbD principles (Service Provider Responsibility, User Empowerment, Transparency/Accountability) are most critical to prioritize in the design, launch, and ongoing operation of this community forum? Justify your choice.

4. **Brainstorm Design/Mitigation Features:**
   - What are 2-3 specific preventative design choices or mitigating features you would prioritize building into this forum? Consider aspects like: content moderation approach (human/AI/hybrid), design of reporting mechanisms (for posts, users, topics), available user controls (blocking users, filtering content), clarity of onboarding information about forum rules, application of trauma-informed principles in reporting flows.

5. **Consider Culture and Implementation:**
   - What is one key cultural element (e.g., required moderator training standards, established cross-team communication protocols for incident response) or practical adoption step needed internally to help ensure this feature is launched and managed safely?

Briefly, how might you ethically involve users (co-design) in shaping the forum's community guidelines or testing safety features before launch?

Process and Share-Out:

Groups are encouraged to actively discuss these points and capture their key ideas concisely. Facilitators circulate to clarify concepts if needed, but primarily allow groups to apply the thinking process themselves. The activity concludes with a brief Share-Out Session, where each group highlights one key risk they identified and one significant proposed safety measure or implementation consideration. This allows for synthesis of common themes and contrasting approaches.

Simulation & Linking to Prototyping Workshops

Facilitator's Note: Use this section to clearly transition from the foundational knowledge learned to the practical application phase in the upcoming workshops. Aim to generate enthusiasm and confirm participants' engagement in the next steps.

The tabletop simulation exercise served as a direct and practical segue to the upcoming hands-on Prototyping Workshops.

We will use the awareness and knowledge you have gained in this training: the principles, processes, design considerations, and implementation factors crucial for effective SbD and apply them in a supported environment. During the upcoming design workshops, participants will apply the learned concepts directly to their own specific products, unique challenges, and distinct user contexts.

Workshop Goals Recap:

- **Workshop 1**: Focus on analyzing your product's specific vulnerabilities and risks (Risk Identification) and brainstorming concrete SbD solutions tailored specifically to your context and user base (Ideation).
- **Workshop 2:** Focus on translating the most promising ideas into tangible and testable prototypes. This includes developing initial plans for conducting ethical user testing of these prototypes and creating preliminary Scaling Roadmaps outlining potential steps for broader implementation if the solutions prove effective.

Participants will utilize resources such as Prototyping Kits (potentially containing user personas relevant to their context, checklists, design templates) and build upon the frameworks introduced during the training. These workshops will be collaborative and hands-on and we will walk away with better SbD features.

Showcase & Recognition: To celebrate this practical work, the program will culminate in a **showcase event.** This will be a fantastic opportunity to share the innovative SbD features and lessons learned with the broader tech community. Furthermore, outstanding designs developed during the workshops will be recognized and awarded prizes.

Summary

This exploration of SbD provided a foundational framework for proactively embedding user safety, rights, and well-being into the core of digital product development and management.

Key Conclusions:

- **SbD as a Paradigm Shift:** SbD represents a necessary move from reactive fixes to proactive prevention, demanding safety considerations be integrated throughout the entire PDLC.
- **Core Principles are Foundational:** The principles of Service Provider Responsibility, User Empowerment and Autonomy, and Transparency and Accountability provide a robust framework for action.
- **Context Matters:** Understanding the specific landscape of digital harms, especially the prevalence and impact of harms targeting users who might experience contextual vulnerability in digital spaces, is crucial for tailoring effective interventions.
- **Culture is Critical:** Sustainable SbD implementation is impossible without visible leadership commitment and the cultivation of a pervasive organizational safety culture where safety is a shared value and responsibility.
- **Process Integration is Key:** Systematically identifying risks (using methods like the 5-step process and threat modelling), assessing impact, and embedding mitigation/prevention controls into standard workflows (PDLC) is essential.
- **Design for Safety and Against Misuse:** Practical design choices must prioritize user safety (e.g., reporting, blocking, privacy controls) while actively considering and mitigating potential misuse (abusability). Trauma-informed design principles are vital for sensitive interactions.
- **Ethical Co-Design Empowers:** Meaningfully and ethically engaging users, particularly those with lived experience of harm, through trauma-informed co-design leads to more relevant, effective, and user-centred safety solutions.

References

1. eSafety Commissioner. esafety.gov.au. [Online].; 2019. Available from: https://www.esafety.gov.au/sites/default/files/2019-10/SBD%20-%20Overview%20May19.pdf.

2. Bundtzen S. Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence. Berlin:; 2023.

3. Communications Authority of Kenya. FIRST QUARTER SECTOR STATISTICS REPORT FOR 2024-2025. Nairobi:; 2025.

4. UN Women, WHO. UN Women. [Online].; 2023. Available from: https://www.unwomen.org/en/digital-library/publications/2023/03/expert-group-meeting-report-technology-facilitated-violence-against-women.

5.IREX. Understanding Threats to Women's Online Safety and Current Pathways to Protection, Prevention, and Redress in Kenya: Analysis of Existing Legislative and Institutional Frameworks and Opportunities for Action. Washington D.C.:; 2025.

6. IREX. irex.org. [Online].; nd. Available from: https://www.irex.org/files/nmwso-program-factsheet.pdf.

7.UNFPA. unfpa.org. [Online].; 2023. Available from: https://www.unfpa.org/sites/default/files/pub-pdf/UNFPA_SafeEthicalGBVTechGuide_Summary_2023.pdf.

8.TSPA. tspa.org. [Online].; n.d. Available from: https://www.tspa.org/curriculum/ts-curriculum/safety-by-design/sbd-what-it-is-why-it-matters/.

9.eSafety Commissioner AU. easfety.gov.au. [Online].; 2024. Available from: https://www.esafety.gov.au/sites/default/files/2024-09/SafetyByDesign-technology-facilitated-gender-based-violence-industry-guide.pdf?v=1726531200021.

10. Australian eSafety Commissioner. easfety.gov.au. [Online].; n.d.. Available from: https://www.esafety.gov.au/industry/safety-by-design.

11. eSafety Commissioner. accc.gov.au. [Online].; 2020. Available from: https://www.accc.gov.au/system/files/Office%20of%20the%20eSafety%20Commissioner%20%28February%202019%29.PDF.

12. Woods L. Online Safety Act Network. [Online].; 2024. Available from: https://www.onlinesafetyact.net/analysis/safety-by-design/.

13. OFCOM. ofcom.org. [Online].; 2025. Available from: https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/ofcom-calls-on-tech-firms-to-make-online-world-safer-for-women-and-girls/.

14. CCDH. Counterhate.com. [Online].; 2023. Available from: https://counterhate.com/blog/star-framework-safety-by-design/.

15.OECD. oecd.org. [Online].; 2024. Available from: https://one.oecd.org/document/DSTI/CDEP(2023)13/FINAL/en/pdf.

16. OECD. Legalinstruments,oecd.org. [Online].; 2025. Available from: https://legalinstruments.oecd.org/public/doc/272/272.en.pdf.

17. unicef. unicef.org. [Online].; 2020. Available from: https://www.unicef.org/innocenti/media/1096/file/%20UNICEF-Global-Insight-DataGov-data-use-brief-2020.pdf.

18. Ofcom. ofcom.org. [Online].; 2025. Available from: 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/main-docs/consultation-document-a-safer-life-online-for-women-and-girls.pdf?v=391803.

19.OWASP. owasp.org. [Online].; nd. Available from: https://owasp.org/www-community/Threat_Modeling_Process.

20. DSIT. gov.uk. [Online].; 2025. Available from: https://assets.publishing.service.gov.uk/media/67a39e2cad556423b636cadd/Platform_design_risk_of_online_violence_against_women_girls_A.pdf.

21. ODPC. odpc.go.ke. [Online].; 2019. Available from: https://www.odpc.go.ke/wp-content/uploads/2024/02/TheDataProtectionAct__No24of2019.pdf.

22. NTIA. Industry's Role in Promoting Kids' Online Health, Safety, and Privacy: Recommended Practices for Industry. Washington D.C.:; 2024.

23. TSPA. Trust and Safety Professionals Association. [Online].; n.d. Available from: https://www.tspa.org/curriculum/ts-curriculum/safety-by-design/implementing-safety-by-design/.

24. Ilana Berger TT. Active Fence. [Online].; 2024. Available from: https://www.activefence.com/what-is-trust-and-safety/.

25. ActiveFence. activefence,com. [Online].; 2022. Available from: https://www.activefence.com/blog/safety-by-design/.

26. OAIC. oaic.gov.au. [Online].; 2025. Available from: https://www.oaic.gov.au/news/blog/a-safer-internet-doesnt-need-to-compromise-privacy.

27. CRS Reports. Social Media Algorithms: Content Recommendation, Moderation, and Congressional Considerations. [Online].; 2023 [cited 2025 May. Available from: https://www.everycrsreport.com/reports/IF12462.html.

28. Lumenalta. The use of artificial intelligence (AI) brings immense potential to transform industries, drive innovation, and solve complex problems. [Online].; 2024 [cited 2025 May. Available from: https://lumenalta.com/insights/ethical-considerations-of-ai.

29. Google. imda.gov.sg. [Online].; 2024. Available from: https://www.imda.gov.sg/-/media/imda/files/regulations-and-licensing/regulations/online-safety/youtube-2024-annual-online-safety-report.pdf.

30. World Economic Forum. weforum.org. [Online].; 2025. Available from: https://www.weforum.org/stories/2025/01/tackling-emerging-harms-create-safer-digital-world-2025/.

30. World Economic Forum. weforum.org. [Online].; 2025. Available from: https://www.weforum.org/stories/2025/01/tackling-emerging-harms-create-safer-digital-world-2025/.

31. World Economic Forum. weforum.org. [Online].; 2025. Available from: https://www.weforum.org/stories/2025/01/tackling-emerging-harms-create-safer-digital-world-2025/.

32. Lewington R. Thriving Games. [Online].; 2022. Available from: https://thrivingingames.org/wp-content/uploads/2022/06/FPA-Being-Targeted-about-Content-Moderation.pdf.

33. 5RightsFoundation. 5rightsfoundation.com. [Online].; 2021. Available from: https://5rightsfoundation.com/wp-content/uploads/2021/09/Pathways-how-digital-design-puts-children-at-risk.pdf.

34. NSPCC. Is Roblox safe for my child? [Online].; 2022 [cited 2025 May 5. Available from: https://www.nspcc.org.uk/keeping-children-safe/online-safety/online-safety-blog/roblox/.

35. CWES. CWES.org. [Online].; 2022. Available from: https://cwes.org.au/wp-content/uploads/2022/11/CWES_DesigntoDisrupt_1_Banking.pdf.

36. Parliament of Australia. aph.gov.au. [Online].; n.d.

37. Bygrave LA. Security by Design: Aspirations and Realities in a Regulatory Context. Oslo Law Review. 2022 June 7: p. 126-177.

38. World Economic Forum. www3.weforum.org. [Online].; 2023. Available from: https://www3.weforum.org/docs/WEF_Global_Charter_of_Principles_for_Digital_Safety_2023.pdf.

39. Perrino J. brookings.edu. [Online].; 2022. Available from: https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/.

40. Hinduja S. cyberbullying.org. [Online].; 2025. Available from: https://cyberbullying.org/empowering-protecting-youth-online-legislation-hinduja-lalani-final.pdf.

41. TechEHS. techehs.com. [Online].; nd. Available from: https://techehs.com/blog/developing-a-culture-of-safety-through-training.

42. OECD. oecd.org. [Online].; 2025. Available from: https://www.oecd.org/en/blogs/2025/03/10-steps-for-policymakers-to-advance-immersive-technologies.html.

43. Telecom Review. telecomreview.com. [Online].; 2024. Available from: https://www.telecomreview.com/articles/reports-and-coverage/8475-designing-child-safe-ai-balancing-innovation-and-digital-safety.

44. Tremosa L. interaction-design.org. [Online].; 2025. Available from: https://www.interaction-design.org/literature/article/beyond-ar-vs-vr-what-is-the-difference-between-ar-vs-mr-vs-vr-vs-xr?srsltid=AfmBOoqs_tIIIWO945PiVqIVBgPPm2pSBFPEKVFgcRlA_1hsXI2Xx3T7.

45. UCLA. Institute of Technology, law and policy. [Online].; 2024. Available from: https://itlp.law.ucla.edu/wp-content/uploads/2024/09/UCLA_ITLP_Governing_XR.pdf.

46. Xiao Li WY. A Critical Review of Virtual and Augmented Reality (VR/AR) Application in Construction Safety. Automation in Construction. 2017;: 86-112.

47. XRSI. Virtual Worlds Real Risks and Challenges. San Francisco:; 2021.

48. GIFIS. Industrial Safety Initiative. [Online].; 2024. Available from: https://www.industrialsafetyinitiative.com/storage/GIFISManifesto_2025.pdf.

49. Grant JI. esafety.gove.au. [Online].; 2019. Available from: https://www.esafety.gov.au/newsroom/blogs/when-smart-is-not-necessarily-safe-the-rise-of-connected-devices-extending-domestic-violence.

50. Design N. Medium. [Online].; 2024. Available from: https://medium.com/@nicoledesign/trauma-informed-design-within-the-digital-product-ui-ux-world-5535d3ed2115.

51. Social Development Direct. sddirect.org.uk. [Online].; 2023. Available from: https://www.sddirect.org.uk/resource/technology-facilitated-gender-based-violence-preliminary-landscape-analysis.

52. SAMHSA. samhsa.gov. [Online].; 2024. Available from: https://www.samhsa.gov/mental-health/trauma-violence/trauma-informed-approaches-programs.

53. Global Platform for Child Explotitation Policy. globalchildexploitationpolicy.org. [Online].; n.d. Available from: http://www.globalchildexploitationpolicy.org/policy-advocacy/safety-by-design.

54. Benjamin D. Trum ea. Safety-by-design and engineered nanomaterials: the need to move. Environment Systems and Decisions. 2023 August 26: p. 178-188.

55. KIDS ONLINE HEALTH AND SAFETY TASK FORCE. Online Health and Safety for CHildren and Youth: Best practices for families and guidance for industry. Washington:; 2022.

56. Woodhouse J. Uk Parliament House of Commons. [Online].; 2025. Available from: https://commonslibrary.parliament.uk/research-briefings/cdp-2025-0043/.

Annex 1: List of Terms and Definitions

- **Abusability Testing**: A practical method focused on deliberately thinking like an adversary to explore how a specific product feature, workflow, or user interaction could be exploited or abused in unintended ways that lead to safety harms.
- **Child Rights by Design (CRbD):** An approach that integrates principles from the UN Convention on the Rights of the Child (CRC) into the product lifecycle for digital services, especially those used by children.
- **Co-Design:** A participatory approach that treats users and community members as active partners and experts in their own lived experience, collaborating directly with designers and developers throughout the design lifecycle.
- **Community Notes (X/Twitter):** A system allowing eligible users to collaboratively add context notes to posts they believe are misleading, representing a decentralized approach to adding context.
- **Content Moderation:** The process of reviewing user-generated content against provider policies and taking action on content that violates those policies (e.g., removal, labeling).
- **Cyberflashing:** The unsolicited sending of sexual messages or intimate images, often via direct messaging.
- **Cyberstalking:** The persistent use of technologies such as GPS tracking, monitoring of social media and other digital activity, and the use of spyware to monitor, track, harass, and intimidate an individual.
- **Dark Patterns:** Deceptive design elements in user interfaces that lead to the manipulation of users into making unintended choices. Examples include hard-to-cancel subscriptions, pre-checked consent boxes, and confirm shaming.
- **Data Minimization**: A principle requiring the collection of only the personal data strictly necessary for achieving a specific, stated purpose.
- **Digital Personas:** Composite profiles or typologies of potential users, including their needs, motivations, vulnerabilities, and digital engagement patterns, used to inform user-centric design processes.
- **Doxxing:** The malicious act of openly publicizing an individual's private or personally identifying information without their consent, often with the intention to incite harassment or harm.
- **Duty of Care:** An ethical and (increasingly) legal responsibility of technology providers to protect users from harm that may arise from the use of their products or services.
- **End-to-End Encryption (E2EE):** A security method where messages are encrypted by the sender and decrypted only by the intended recipient, preventing even the provider from accessing the content.
- **Gendered Disinformation**: The intentional creation and spread of false or misleading narratives specifically designed to target women or girls, often to discredit them or discourage their participation in public life.
- **Hashing Technology (for Safety):** A process that converts data (like an image file) into a unique, fixed-size string of characters (the "hash"). In safety applications, if a known harmful image has its hash stored, digital product and services can block uploads matching that hash without needing to see the image itself, thus preventing its spread.
- **Image-Based Abuse (IBA)**: Encompasses the creation, manipulation, non-consensual distribution, or threat of distribution of intimate or sexual images and videos of an individual without their explicit consent.
- **Internet of Things (IoT):** The network of everyday physical objects embedded with sensors, software, and other technologies enabling them to connect to the internet and exchange data with other devices and systems.
- **Metaverse:** A concept for a persistent, shared, 3D virtual space or network of virtual worlds where users, often represented by avatars, can interact with each other and digital environments in real-time for socializing, work, commerce, and entertainment.

- **Non-Consensual Intimate Imagery (NCII):** Intimate or sexual images and videos created, manipulated, distributed, or threatened with distribution without the explicit consent of the person depicted.
- **Product Development Lifecycle (PDLC):** The sequence of stages a product goes through, typically including Ideation/Concept, Design/Planning, Development/Implementation, Testing/Quality Assurance, Launch/Deployment, and ongoing Maintenance/Monitoring/Iteration.
- **Privacy Impact Assessment (PIA):** A systematic process to identify and assess the privacy risks arising from the processing of personal data and to develop appropriate mitigation measures.
- **Red Teaming (Safety-Oriented):** Practical exercises where internal or external experts act as adversaries to proactively identify vulnerabilities, potential avenues for abuse, and ways a product or feature can be weaponized or misused before launch.
- **Safety-by-Design (SbD):** A proactive approach that puts user safety and rights at the center of the design and development of digital products and services, integrating safety features and considerations throughout the product lifecycle to prevent and mitigate harms before they occur.
- **Safety Culture (Organizational):** Shared values, beliefs, common attitudes, and consistent behaviors concerning user safety that exist at every level of an organization, where user well-being is actively and consistently prioritized.
- **STRIDE:** A mnemonic for a threat modelling framework categorizing threats as: Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege.
- **Technology-Facilitated Gender-Based Violence (TFGBV):** Any act rooted in gender inequality that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms.
- **Threat Modelling:** A structured approach to identify, quantify, and address security risks associated with a product or tool, often by adopting the perspective of an attacker.
- **Trauma-Informed Design:** An approach to designing products, features, interfaces, and communication styles that actively avoids causing further harm or re-traumatization to users who may have diverse experiences with trauma, while promoting user agency, healing, and a sense of safety.
- **Virtual Reality (VR) Headsets:** Devices that fully replace a user's view of the real world with a computer-generated, immersive 3D environment using stereoscopic displays and head-tracking technology, differing from audio headphones or AR glasses which overlay information or only provide sound.

SAFETY BY DESIGN

ADDENDUM

1

SAFETY-BY-DESIGN IN AI-POWERED TOOLS

SAFETY by Design reflects the collaboration and contribution of many people and organizations engaged in preventing, responding to, and mitigating online harms. All sources have been cited.

Prepared by IREX with the safety-by-design expertise of Eugene Odanga Masinde, reviews and feedback from Tech Innovators Network (THiNK), Development Gateway, and professional graphic design of Tamar Gabisonia. The team would also like to thank the experts at Australia's eSafety Commission for their invaluable guidance and feedback.

Date of publication: February 2026

LLM	Large Language Model; an AI system trained to understand and generate human-like language.
Chatbot	A tool or system that 'talks' with users, answering questions or completing tasks using written or spoken language.
Human-in-the-loop	A process where humans stay involved to review, guide, or correct what an AI system produces.
API	Application Programming Interface; a way for different software systems to share information or functions.
RACI	Responsible, Accountable, Consulted, Informed; a tool, often a spreadsheet, that clarifies project roles and tasks.
UX	User Experience; often referring to the process of creating user-friendly interfaces and/or the team in charge of these tasks.

ADDENDUM 1

Safety-by-Design in AI-Powered Tools

Learning Objectives

Upon completion of this module, participants should be able to:

1. Articulate broad AI safety principles
2. Differentiate between safety, security, and ethics in AI systems
3. Understand that different types of AI applications require nuanced approaches
4. Identify common risks and harms associated with integration of AI chatbots into products and services
5. Map SbD principles across all stages of the AI chatbot lifecycle
6. Be able to apply proactive risk prevention and inclusive strategies within their company/organization

Materials Needed

- Participant handouts
- Markers
- Flipcharts

Time Allotted

2 hours

Defining Artificial Intelligence

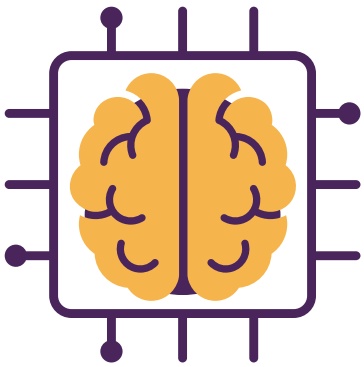
Artificial Intelligence (AI) is technology that works by learning from large amounts of data and using that knowledge to provide responses, predictions, or actions in real time, mimicking human intelligence. (12)

Is AI	Is *NOT* AI
Uses machine learning models (e.g., LLMs, classifiers, recommendation models)	Uses rules, decision trees, or if-else logic
Generates content, responses or actions using learned patterns	Generates responses from fixed scripts or menus
Can misinterpret, hallucinate, or make unsafe inferences	Mostly predictable unless logic is faulty
Improves over time using new data	Behaves the same unless manually updated

Example:

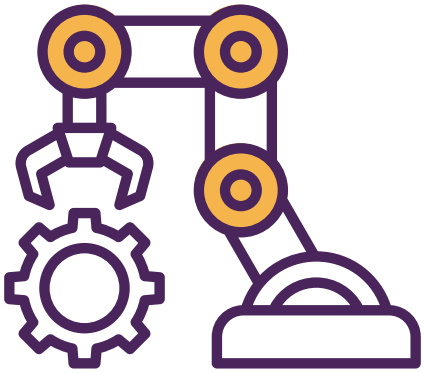
A chatbot user types: "I have a question about my mental health."

AI Chatbot Response



"Hi — I'm glad you reached out. I'm here to listen. What's been going on for you, or what question do you have about your mental health? You can share as much or as little as you feel comfortable with."

Non-AI Chatbot Response



"Please select an option:
1. Mental health resources
2. Speak to a live specialist"

AI is a uniquely powerful form of technology because it can learn from data, adapt to new situations, create new content, and deliver personalized experiences - helping users make decisions faster, access tailored information, and automate repetitive or complex tasks. (12) These benefits make AI transformative across sectors like health, finance, and education. However, the same capabilities that enable personalization and scale can also amplify risks if safety isn't built in from the start. Without safety-by design (SbD), AI systems can cause harm to individuals and at scale, mislead users, limit access to information and options for specific users, spread misinformation, reinforce bias, invade privacy, or manipulate user behavior. Embedding safety principles early ensures that AI delivers its benefits responsibly while minimizing harm.

Why Safety by Design Matters in AI

As with all technology, AI systems, applications, and elements must be safe, secure, and ethical.

- Safe = Prevent harm from unintended behavior
- Secure = Protect against external threats and misuse
- Ethical = Ensure moral and societal responsibility

An SbD approach ensures all these key pillars are in place and helps make AI tools reliable, ethical, and structured in ways that do not cause harm to individuals, communities, or society. It involves designing, testing, and iterating AI to prevent risks while promoting transparency, accountability, and user trust. As with all technology, SbD in AI is about embedding safeguards throughout the lifecycle of the system so it delivers benefits without unintended or harmful consequences. (20, 24)

Impact of Unsafe AI

Poorly developed AI systems don't just harm individual users - they can create ripple effects across society, economies, and organizations - even the ones that created them.

Here are just some of the documented statistics:

- In January 2026, a report by Centre for Information Resilience spotlights how unrestricted AI image generation on Grok AI enabled widespread production of non-consensual, sexually explicit, and abusive imagery, disproportionately targeting women and children – undermining trust and online safety. Out of 1,625 prompts for Grok image generation, 72% targeted women and 98% of those were sexualized requests. (4)

- In 2024, the click-through rate for AI-generated phishing content was 54%, versus 12% for traditional phishing methods. (17)

- A Psychiatric Times review found that at least 27 AI chatbots (including ChatGPT, Character.AI, Replika, Woebot, and others) have been associated with 10 types of adverse mental health outcomes, such as self-harm, delusions, psychosis, sexual harassment, and suicide encouragement. (23)

- Independent academic audits and reviews continue to find that AI-assisted hiring tools can produce discriminatory outcomes across race and gender. For example, a University of Washington team simulated large-scale résumé screening with modern LLM-based systems and observed systematic racial and gender bias in ranking outcomes, including frequent preference for White-associated names and consistent disadvantage for Black male–associated names across occupations. (27)

Let's review and prepare to discuss some potential broader impacts:

Organizational Risks

- **Legal Liability:** Organizations deploying unsafe AI may face lawsuits for privacy breaches, discrimination, or harm caused by biased outputs. Non-adherence to data protection laws and AI ethics standards can also result in regulatory penalties and reputational damage.
- **Financial Losses:** Unsafe deployments can lead to costly recalls, loss of consumer trust, and decreased adoption of products.

Societal Consequences

- **Misinformation:** AI-driven bots can create and spread false narratives at scale, influencing public opinion and undermining trust in institutions.
- **Polarization:** Algorithmic amplification of divisive content can deepen social and political divides, creating echo chambers and reducing constructive dialogue.
- **Bias and Discrimination:** AI systems trained on biased data can perpetuate or amplify discrimination in areas like banking and law enforcement.
- **Loss of Privacy:** AI can facilitate mass surveillance and data collection.
- **Security Risks:** AI can be employed to facilitate cyberattacks, fraud, and other illegal and/or dangerous activities.

Discussion:

> Can you think of a community-level or societal impact of unsafe AI that's happened or is currently happening in your community?

AI-related risks add to existing risks associated with digital products and services: integrating AI into digital products and services can trigger new risks beyond those already driven by factors like data collection and use practices, user age, integration of third-party providers, user engagement formats, sector of product/service, and others. Those risks are covered in other modules of the SbD curriculum. This module dives deeper into understanding vulnerabilities specific to AI, using AI chatbots as a case study due to their common use and broader familiarity. There are also multiple other applications of AI in the tech sector, such as predictive analytics, computer vision, natural language processing, generative AI, and process automation, which all have their own specific risks and mitigation measures.

Current Safeguards

As this new frontier emerges, the world is catching up with how to keep these new tools and resources safe for everyone. The processes, standards and guardrails that help ensure AI systems and tools are safe are often referred to as AI governance. At its core, AI governance aligns closely with the three pillars discussed earlier and ensured through a Safety-by-Design approach:

- Safe - preventing harm from unintended behavior
- Secure - protecting against external threats and misuse
- Ethical - upholding moral and societal responsibility

Safety and Security Practices

At an organizational level, this work includes broader oversight and control mechanisms integrated through the lifecycle of an AI system. (13) These examples can apply to all forms of AI, and are crucial as we explore the case of AI chatbots in particular.

- **Establish Clear Accountability:** Define roles and responsibilities for AI oversight across technical, legal, and operational teams.
- **Keep Records:** Maintain easily accessible logs and audit trails for accountability and to facilitate reviews of AI systems' decisions and behaviors.
- **Human Oversight:** Maintain human-in-the-loop mechanisms for high-risk decisions to prevent automation errors.
- **Incident Response Protocols:** Develop clear escalation pathways for addressing AI failures or harmful outputs.

Ethical Codes of Practice

In parallel, global regulatory bodies and governments are increasingly introducing both binding and non-binding Codes of Ethics and governance frameworks to guide responsible AI practices and mitigate risks. UNESCO, the African Union (AU), European Union (EU) and Organization for Economic Co-operation and Development (OECD) are four well-known examples.

UNESCO Recommendation on the Ethics of Artificial Intelligence:
A global standard for AI ethics, agreed upon by all 193 UN Member States to guide the responsible development and use of AI worldwide. Its core purpose is to ensure that AI systems protect human rights, dignity, equality, and the environment, while supporting sustainable development and social good. (24)

AU Continental Artificial Intelligence Strategy:
Africa's first unified framework for responsible AI development, endorsed in July 2024. It promotes an Africa-centric, ethical, and inclusive approach focused on five priorities: maximizing AI's benefits, strengthening data and digital infrastructure, building skills, minimizing risks through rights-based safeguards, and advancing regional cooperation. (1)

EU AI Act:
The EU AI Act is the first comprehensive legal framework for AI worldwide, introducing a risk-based approach with four levels of risk (unacceptable, high, limited, minimal) and strict compliance requirements for high-risk systems, including transparency and human oversight. Non-compliance can result in fines up to €35 million or 7% of global turnover. (5, 6, 7)

OECD AI Principles:
Adopted by 47 countries, these principles promote trustworthy AI through five pillars: inclusive growth, respect for human rights, transparency, robustness, and accountability. They were updated in 2024 to address generative AI risks, including privacy, safety, and information integrity. (20)

Emerging National Codes of Practice

Global AI Law Tracker:
Countries worldwide are rolling out national AI strategies, ethics policies, and regulatory frameworks to balance innovation with risk management. (11)

CASE STUDY: Kenya's National AI Strategy

Launched by the Ministry of Information and Communications Technology (ICT) in 2024, Kenya's National AI Strategy sets out a vision to position the country as a regional leader in AI research, innovation, and commercialization. It emphasizes data sovereignty, leveraging local talent and datasets, and transforming priority sectors such as agriculture, healthcare, and public services. The strategy builds on Kenya's existing legal frameworks—including the Data Protection Act (2019) and ICT Policy (2019)—and calls for coordinated governance, stronger digital infrastructure, and partnerships across government, industry, academia, and civil society. (18) Complementing this broader strategy is the AI Code of Practice KS 3007:2024, developed by the Kenya Bureau of Standards (KEBS) in 2024 as dedicated technical and governance guidance for AI systems. It outlines principles for responsible AI deployment, including transparency, human oversight, data protection compliance, consumer safeguards, and accountability. (14)

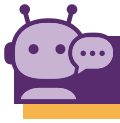
Discussion: AI Safety Principles

Format: Small group discussion and presentation

Participants work in small groups (3 – 4 people) to share, discuss, and note responses to the questions below. They are prepared to present briefly to the rest of the group at the end of their conversation.

Discussion questions: "What AI Safety principles, policies, and guidelines affect your work?
Are there any additional AI ethics considerations missing from the current guidelines and policies that you would like to see in your sector?"

Facilitator note: Facilitator elevates and highlights the most reported types of harms after all groups have shared, and mentions that they will be revisited as the training continues.



CASE STUDY: AI Chatbots

AI chatbots are unique because they interact directly with users—often in real time and sometimes in sensitive contexts like health, finance, or emotional support. This immediacy and personalization make them powerful tools but also introduce risks that go beyond traditional software. Unlike static systems, chatbots can directly influence decisions, collect personal data, and shape user behavior through conversation. These systems can reach users immediately and easier than ever before with the rise in AI chatbot popularity. Misuse or dangerous design can spread misinformation, reinforce harmful biases, compromise privacy, or cause emotional and mental harm.

AI chatbots have become a popular and convenient addition to many digital products because they provide instant, 24/7 assistance, reduce operational costs, and enhance user engagement. They streamline customer support by answering common queries, guide users through complex processes, and personalize interactions based on user data.

For example, in healthcare, chatbots help patients schedule appointments and access basic medical advice; in education, they assist learners with tutoring and answering questions; and in e-commerce, they recommend products and handle order tracking. Their ability to deliver quick, automated responses makes them an attractive solution for businesses seeking efficiency and improved user experience.

Risks Associated with AI Chatbots

AI chatbots can pose significant security, safety, and ethical risks. From a security perspective, they can be exploited through adversarial attacks, data leaks, or prompt injections, exposing sensitive user information and enabling phishing scams and hacking. In terms of safety, poorly designed or unmoderated chatbots may provide inaccurate advice, encourage harmful behaviors, or fail in critical contexts like healthcare, leading to real-world harm. On the ethical front, chatbots can perpetuate bias, manipulate emotions, and lack transparency in decision-making, raising concerns about fairness, accountability, and user trust.

Risk	Illustrative Examples
Manipulation into overspending	• A retail chatbot suggests premium products and uses persuasive phrases like "Most customers upgrade for better results." • A banking chatbot offers instant credit or loan options during a conversation about budgeting, predicting acceptance based on past behavior. • A gaming chatbot encourages in-app purchases by highlighting "exclusive offers" and creating urgency with countdown timers.
Use of data (location) without consent	• A delivery chatbot asks for your address and stores it without clear consent or retention limits. • A ride-hailing chatbot infers your home location from repeated pickup points and shares it with third-party advertisers. • A social chatbot casually asks "Where are you chatting from?" and uses that data for targeted ads without disclosure.
Bias (Unintentional and Intentional)	• A hiring-assistant chatbot ranks candidates lower if they attended women's colleges or took career breaks, reflecting biased training data rather than actual qualifications. • A medical triage chatbot underestimates pain levels or risk scores for certain ethnic groups because its model was trained on non-representative datasets. **CASE EXAMPLE** *KENYA*: A Kenyan customer-support chatbot misinterprets user input in Swahili or Sheng as "non-compliant," leading to more frequent non-response or account flags for non-English-speaking users.
Reality confusion, reinforcement of delusional beliefs or overreliance on digital tools for emotional or mental support	• A mental health chatbot markets itself as a "virtual therapist" and encourages users to rely on it for emotional support without disclaimers or referrals to real-world help. • A companion chatbot uses anthropomorphic language (I'll always be here for you") and creates dependency. • A chatbot in a discussion forum repeatedly validates conspiracy theories or pseudoscientific claims instead of providing balanced information.

Activity: AI Risk Poll

Format: Small group discussion or digital poll

Participants shout out answers or provide them via digital tool such as a poll or word cloud.

Discussion questions: "What are the top three perceived risks associated with AI chatbots in your products and services?"

Facilitator note: Facilitator calls on participants who don't speak up immediately. After responses, discuss and review the list below.

1. **Privacy and Data Security** – AI chatbots often collect personal data, which can be misused or exposed in data breaches.
2. **Misinformation** – AI chatbots can provide inaccurate or outdated information, especially if they rely on incomplete or biased training data. In sectors like health or finance, this can lead to harmful decisions.
3. **Bias and Discrimination** - AI chatbots can reflect biases present in their training data, leading to discriminatory responses.
4. **Over-Reliance** - Users may trust chatbots too much, assuming they are always correct. This can reduce critical thinking and lead to poor decision-making.
5. **Lack of Transparency** - Many chatbots do not clearly explain how they work or what their limitations are. Users, especially youth and children, may not realize they are interacting with an AI rather than a human.
6. **Security Vulnerabilities** – AI chatbots can be exploited by attackers for phishing, hacking, or spreading malware. Poorly secured systems could become entry points for cyberattacks.
7. **Emotional Manipulation** – AI chatbots designed for engagement can exploit psychological vulnerabilities, leading to addictive use or undue influence.
8. **Regulatory and Compliance Risks** - In sectors like healthcare or finance, AI chatbots may inadvertently violate laws or regulations if not properly designed/updated.

Making AI Chatbots More Secure, Safe, and Ethical through SbD

To enhance safety and security, Safety-by-Design principles emphasize embedding robust protections into the chatbot's architecture from the start. This includes implementing strong authentication and encryption, securing APIs, and using adversarial testing to identify vulnerabilities before deployment. Continuous monitoring for prompt injections, data poisoning, and model theft is essential, along with clear data governance policies to prevent unauthorized access or leaks.

Ethically, SbD requires proactive measures to minimize harm and uphold fairness. Developers should integrate bias detection tools, enforce transparency through explainable AI, and set strict guardrails to prevent, detect, and respond to harmful or manipulative outputs. Human-in-the-loop review for high-risk interactions, real-time threat intelligence, clear disclaimers about limitations, and inclusive training datasets help ensure equitable outcomes. Additionally, aligning chatbot behavior with ethical frameworks —such as OECD AI Principles or UNESCO guidelines—builds trust and accountability, adapts to evolving threats while reducing unintended consequences.

Let's explore concrete features that make AI chatbots safer, more inclusive, and more resilient to misuse.

Activity: Features That Strengthen Safety in AI Chatbots

Format: Small group discussion and presentation.

Participants' task is to examine the list of safety-strengthening features and identify relationships between the safety features, the type of chatbots that might benefit from them, and the risks that they might address.

Instructions:
Review the list of Safety-by-Design features on the next page. In the worksheet on page XX, fill in the missing cells in each row of the table based on what information is included in the other columns. For example, if row 1 already has "Human content moderation (e.g. human review of posts/comments)" listed in the Safety-by-Design column, participants should fill out the empty cell in row 1 under the column "AI Tool Type" with any kind of AI tool that could benefit from that feature (i.e. A health advice chatbot). If the cell in row 1 under the column "Risks Addressed" is also blank, participants should write in that cell which risks the listed safety-by-design feature could help prevent or mitigate (i.e. Biased information that leads to poor decisions") Answers may vary.

Guiding Questions:
Which user groups are most at risk?
Which harms emerge in high-stakes contexts?
Can safety features address multiple harms?

Example:

AI Tool Type	Safety-by-Design Feature	Risk Addressed
Example Response: Health Advice Chatbot	Example Response: Human content moderation (e.g. human review of posts/comments)	**Inaccurate, biased or low quality info leads to poor or dangerous decisions**
[participants fill in what type of AI tool might need this feature]	**Transparency about automation (e.g. clear labeling of bots)**	*[participants fill in what type of risk this feature addresses that could also occur in the type of AI tool they list]*
[participants fill in what type of AI tool would use the feature they list and have the risk]	*[participants fill out what feature might address the risk]*	**Data leaks, security breaches, hacking**

Activity: Features That Strengthen Safety in AI Chatbots

Illustrative List of Safety-by-Design Features

Access to support resources	Age verification	Age-appropriate design elements
Age-appropriate language for safety and security policies and settings	Alternative text for images and graphics	Bias audits and fairness testing
Child-friendly content filters	Child-specific privacy settings	Content filters for sensitive topics
Cross-border/overseas disclosure	Data minimization	Device/app access permissions
Educational onboarding for children	Ethical AI guidelines adherence	Explainability features
Feedback loops for harmful behavior detection	Feedback mechanisms for AI errors	Font size adjustment or zoom functionality
Granular consent management	Guardian dashboards	High contrast mode or customizable color schemes
Human-in-the-loop review	Input and output validation	Limited scope of automation
Multilingual support	Non-addictive design features	Offline access for users with limited connectivity
Periodic user education	Positive behavioral nudges	Privacy-enhancing technologies
Proactive data protection prompt	Quick exit option	Regular privacy audits
Reporting and escalation pathways	Restricted personalization and profiling of children	Safe fallback options
Screening for predatory AI bots	Secure biometric handling	Security controls and manuals easily accessible and intuitive
Simplified language or easy-read mode	Text-to-speech or screen reader compatibility	Third-party data sharing disclosures
Transparency about automation	Transparent data use policies	Transparent reporting on moderation practices
User consent for data collection	User control over notifications	User data control tools
Verified user badges	Vetting of third-party entities	Voice command or speech recognition

Activity: Features That Strengthen Safety in AI Chatbots

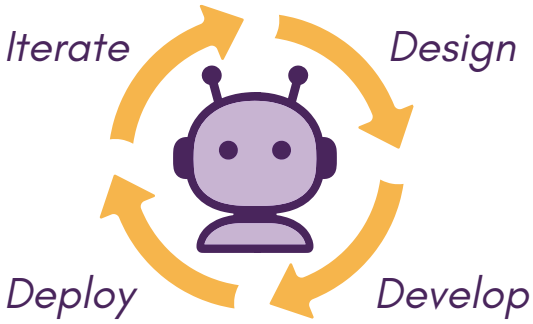
Complete the Table

AI Chatbot Tool	*Safety-by-Design Feature*	*Risk Addressed*
Example Response: Health Advice Chatbot	*Example Response: Human content moderation (e.g. human review of posts/comments)*	**Inaccurate, biased or low quality info leads to poor or dangerous decisions**
	Transparency about automation (e.g. clear labeling of bots)	
		Data leaks, security breaches, hacking
		Exclusion of users with disabilities
		Reality confusion and overdependence on AI
Workplace HR Chatbot		
	Multilingual Support	

Embedding Safety in the Lifecycle of a Chatbot

It is important to have a practical roadmap for embedding safety-by-design principles throughout the lifecycle of an AI or digital-product initiative. The guidance below outlines what cross-functional teams should do at each phase from Design, Development, Deployment, to Iteration, and clarifies the specific responsibilities of key roles involved in building and maintaining safe, trustworthy systems.

In review of the steps that follow, focus on recognizing how early risk mapping, inclusive design practices, robust testing, and continuous monitoring work together to reduce harm, strengthen user trust, and align products with evolving global standards for responsible technology. This framework is intended to support decision-making, improve collaboration across disciplines, and help you operationalize safety in day-to-day practice.



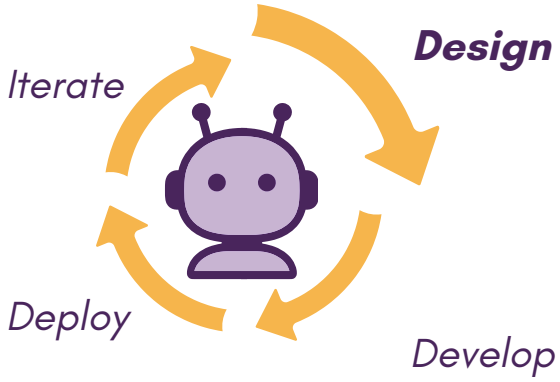
Design Phase

Risks and harms are mapped early, involve diverse users, and define safety goals and standards.

What Teams do at this Stage

- **Map threats, risks & harms** early (privacy, bias, misinformation, prompt injection, sensitive-context misuse). Create a risk register with the likelihood and severity of each harm. (18)
- **Set governance & oversight** (who decides, who signs-off). Establish roles, RACI, escalation paths and human-in-the-loop checkpoints for high-risk interactions; align with AI governance as "processes, standards, guardrails/ multi-layered safety architecture" and broad oversight & control throughout the lifecycle. (13)
- **User & stakeholder consultation** (esp. vulnerable users). Incorporate co-design and ethical guardrails consistent with recommendations such as UNESCO's (human-rights, transparency, accountability). (24)
- **Red teaming** or simulating real-world attacks to test defenses and processes for vulnerabilities (14)
- **Human oversight** requirements for high-risk contexts (e.g., health, finance); design interfaces so humans can monitor, interpret, override; avoid over-reliance. (5).

IREX's PRIMA (Predictive Risk and Mitigation Audit) is one example of a risk register that tech teams can use when designing safe AI chatbots. Using a series of questions on a tool's function, design, and target audience, PRIMA helps developers identify and flag potential privacy, safety, and security risks. PRIMA also documents the likelihood of each risk and highlights relevant safety features that could effectively mitigate or prevent harm.



Role-Specific Tasks (Design Phase)

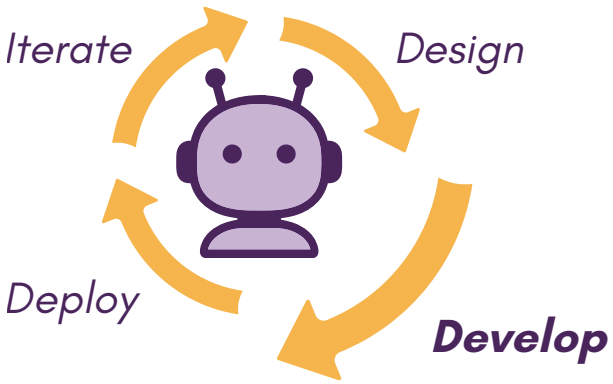
- **Product Manager:** Draft use-case profile aligned to relevant frameworks (purpose, users, context, harms), define KPIs for safety (e.g., reduction in harmful outputs; time-to-mitigation), and set approval gates. (18)
- **UX/Content Design:** Create trauma-informed, transparent onboarding (limitations, quick-exit, reporting) and clear bot disclosure per relevant legal requirements for chatbots.
- **Trust & Safety / Policy:** Specify prohibited content/behaviors, reporting/appeals flows, and enforcement transparency. (18, 24)
- **Security/Privacy Engineer:** Define data-minimization, encryption standards, secure model-update processes, and monitoring for anomalous or abusive inputs. Ensure privacy and security principles are applied from the outset.
- **Legal/Compliance:** Map all applicable legal and standards-based obligations (data protection, consumer-protection, sector-specific rules); ensure that contracts, terms of use, and disclosures reflect model limitations, data uses, and user rights.

Development Phase

Bots are trained on diverse datasets, safety guardrails are built, and bots are tested with adversarial prompts.

What Teams do at this Stage

- **Use diverse and representative training data**, documenting where it comes from, how it was cleaned, and any limitations.
- **Conduct adversarial and robustness testing**, exploring worst-case prompts, language cases, and attempts to bypass safeguards. (21)
- **Document model behavior**, constraints, and limitations to support transparency and responsible release.

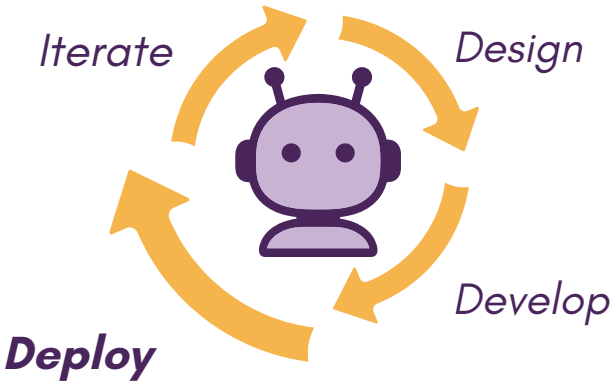


Role-Specific Tasks

- **Product Manager:** Track progress toward safety goals and ensure mitigating actions are implemented before launch.
- **UX/Content Design:** Test safety-related UX flows, including refusal responses, escalation messaging, and multilingual behavior.
- **Legal/Compliance:** Ensure datasets, outputs, and system behavior align with legal and ethical expectations across the lifecycle.

Deployment Phase

Controlled rollouts (pilots) executed, with real-time monitoring and clear user onboarding with safety tips.



What Teams do at this Stage

- **Conduct a controlled pilot**, monitoring live interactions for safety signals, performance issues, or unexpected harmful outputs.
- **Perform operational readiness checks**, including human-review workflows, fallback responses, and alerting systems.
- **Document deployment decisions**, residual risks, and any remaining mitigations needed.

Role-Specific Tasks

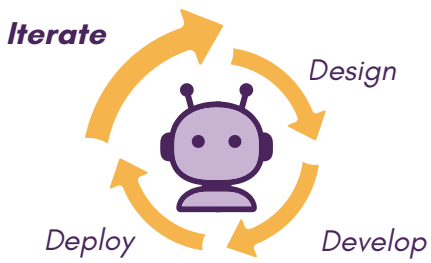
- **Product Manager:** Oversee pilot signals, evaluate whether safety goals are met, make rollout decisions.
- **UX/Content Design:** Ensure onboarding, consent flows, and language-specific messaging are clear, inclusive, and accessible.
- **Legal/Compliance:** Verify that deployment adheres to legal obligations and that user-facing documentation accurately reflects system behavior.

Iteration Phase

Continuous audits are performed, team responds to feedback and aligns with evolving AI Code of Practice.

What Teams do at this Stage

- **Review user feedback** and incidents, adjusting safety measures based on emerging issues.
- **Update guardrails**, training data, and logic to reflect new risks, evolving language, and changing user behavior.
- **Maintain alignment with evolving internal or sectoral AI codes of practice**, applying continuous improvement principles seen in global governance discussions.



Role-Specific Tasks

- **Product Manager:** Lead regular risk and performance reviews and update the product roadmap accordingly.
- **UX/Content Design:** Refresh UX patterns, disclosures, safety indicators, and crisis-support messaging based on user needs and responsible-AI guidance.
- **Legal/Compliance:** Track evolving regulatory expectations, update documentation, and ensure the system remains aligned with legal and policy requirements.

Activity: Risk Identification Simulation

Format: Gamified small group discussion and presentation. Groups of 5-10.

Participants' task is to examine the list of safety-strengthening features and apply them to real-world AI tools through mapping.

Scenario: Designing an AI-driven health chatbot.

Instructions:

- Draft a concept of a chatbot to address health issues for an audience. Attempt to safeguard against possible harms by listing the safety features.
- After 10-15 minutes for creation, teams swap to 'Red-Team' other teams' tools and identify possible harms (misdiagnosis, misinformation, harassment, privacy breach) and propose mitigations (another 10-15 minutes)
- After both phases are complete, teams share out their chatbot concept and what safety features have been recommended and why those features might be included.

Review and Apply:

Red Teaming is a structured, adversarial testing process in which developers deliberately probe an AI system to uncover flaws, vulnerabilities, and potential harms before real users encounter them. It is a structured effort to identify flaws and vulnerabilities, including harmful or discriminatory outputs, unforeseen behaviors, or misuse risks, by using adversarial methods in a safe environment. In the context of Safety-by-Design, red teams act as "ethical adversaries," stress-testing systems under realistic or worst-case conditions—such as attempts to trigger misinformation, unsafe medical advice, privacy breaches, discriminatory outputs, or misuse scenarios. This method helps expose risks that may be overlooked during normal development and strengthens the product's safety features by informing clearer guardrails, mitigations, and human-oversight mechanisms.

References

1. African Union. (2024). Continental artificial intelligence strategy. https://au.int

2. Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. PACMHCI.

3. Center for AI and Digital Policy. (2024). AI frameworks sourcebook. https://www.caidp.org

4. Centre for Information Resilience. (2026, January 14). "Grok'd": Five emerging lessons on limiting abuse of AI image generation. https://www.info-res.org/cir/articles/grokd-five-emerging-lessons-on-limiting-abuse-of-ai-image-generation/

5. European Commission. (n.d.). Regulatory framework for AI. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

6. European Commission. (2024). AI Act enters into force. https://commission.europa.eu/news-and-media/news/ai-act-enters-force-2024-08-01_en

7. European Commission. (n.d.). Rules for trustworthy artificial intelligence in the EU. https://eur-lex.europa.eu/EN/legal-content/summary/rules-for-trustworthy-artificial-intelligence-in-the-eu.html

8. Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F. J., & Biega, A. J. (2025). Fairness and bias in algorithmic hiring: A multidisciplinary survey. ACM Transactions on Intelligent Systems and Technology.

9. Google. (2024). Responsible generative AI toolkit. https://ai.google/responsible-ai

10. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2020). Ethically aligned design. https://ethicsinaction.ieee.org

11. International Association of Privacy Professionals. (n.d.). Global AI legislation tracker. https://iapp.org/resources/article/global-ai-legislation-tracker/#global-ai-chart

12. IBM. (n.d.). Artificial intelligence: What it is and why it matters. https://www.ibm.com/think/topics/artificial-intelligence

13. IBM. (n.d.). AI governance. https://www.ibm.com/think/topics/ai-governance

14. IBM. (n.d.). What is red teaming? IBM Think. https://www.ibm.com/think/topics/red-teaming

15. Kenya Bureau of Standards. (2024). AI Code of Practice KS 3007:2024. https://unece.org/sites/default/files/2024-08/MOkoth_KEBS_Kenyan_AI_CoP.pdf

16. Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? Management Science, 65(7), 2966–2981.

17. Microsoft. (2025). Microsoft Digital Defense Report 2025. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/bade/documents/products-and-services/en-us/security/Microsoft-Digital-Defense-Report-2025.pdf

18. Ministry of Information and Communications Technology (Kenya). (2024). Kenya National AI Strategy. https://ict.go.ke/node/641

19. National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework

20. Organisation for Economic Co-operation and Development. (2024). OECD AI principles. https://oecd.ai/en/ai-principles

21. OWASP. (n.d.). AI adversarial testing documentation. https://owaspai.org/docs/5_testing/

22. Partnership on AI. (2024). Resource library. https://partnershiponai.org/resources

23. Psychiatric Times. (n.d.). Preliminary report on dangers of AI chatbots. https://www.psychiatrictimes.com/view/preliminary-report-on-dangers-of-ai-chatbots

24. UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

25. U.S. Equal Employment Opportunity Commission. (2023). Select issues: Assessing adverse impact in software, algorithms, and AI used in employment selection procedures under Title VII.

26. U.S. Equal Employment Opportunity Commission & U.S. Department of Justice. (2022). Algorithms, AI, and disability discrimination in hiring: ADA technical assistance.

27. Wilson, K., & Caliskan, A. (2024). Gender, race, and intersectional bias in AI résumé screening via language-model retrieval. AIES 2024.