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

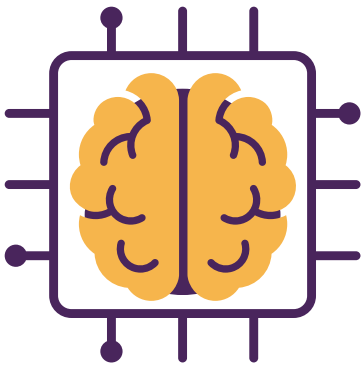
Artificial Intelligence (AI) is technology that works by learning from large amounts of data and using that knowledge to provide responses, predictions, or actions in real time, mimicking human intelligence. (12)

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:

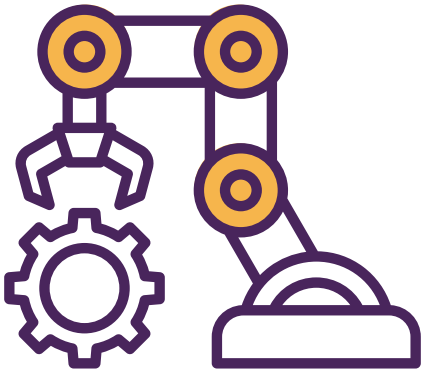
A chatbot user types: "I have a question about my mental health."

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)

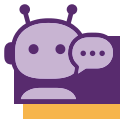
Discussion: AI Safety Principles

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

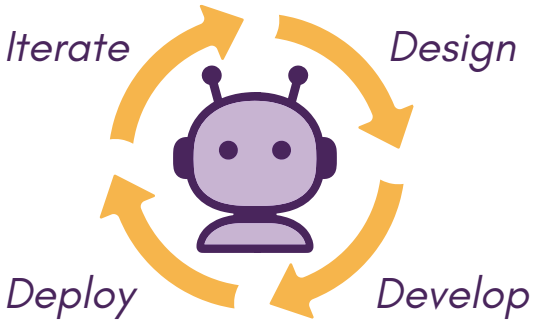
Complete the Table

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.



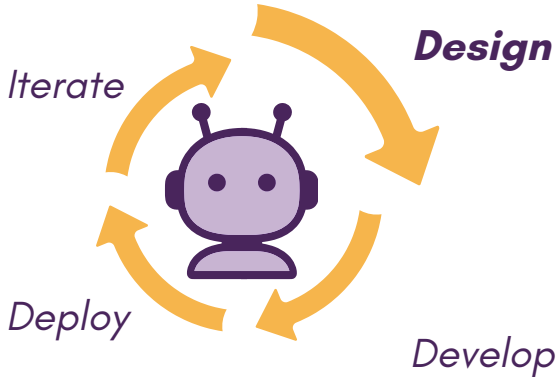
Design Phase

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)

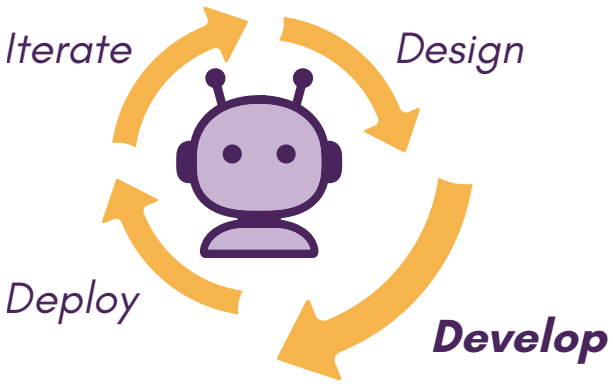
- **Product Manager:** Draft use-case profile aligned to relevant frameworks (purpose, users, context, harms), define KPIs for safety (e.g., reduction in harmful outputs; time-to-mitigation), and set approval gates. (18)
- **UX/Content Design:** Create trauma-informed, transparent onboarding (limitations, 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.

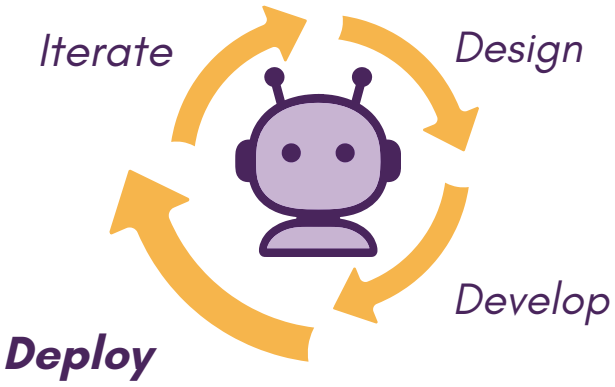


Role-Specific Tasks

- **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

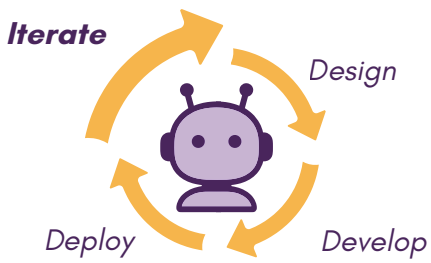
- **Product Manager:** Oversee pilot signals, evaluate whether safety goals are met, make rollout decisions.
- **UX/Content Design:** Ensure onboarding, consent flows, and language-specific messaging are clear, inclusive, and accessible.
- **Legal/Compliance:** Verify that deployment adheres to legal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