AI Safety Playbook

- About This Playbook
- Module 0
  - Safety by Design: Guidance for AI Bot Builders
- Module 1
  - Module 1: Introduction to AI Safety

About This Playbook

> Kenya is rapidly embracing artificial intelligence (AI) as a driver of innovation, economic growth, and improved public service delivery. As AI adoption accelerates across sectors, there is an increasing need to ensure that AI systems are developed and deployed responsibly, ethically, and in ways that protect the rights, safety, and trust of users. This aligns with Kenya's evolving AI governance landscape, including the Data Protection Act, emerging national AI policy initiatives, and international best practices for responsible AI.
>
> The **Safety by Design Playbook** has been developed by **Tech Innovators Network (THiNK)** as a practical guide for AI innovators, developers, startups, researchers, and organizations building conversational AI systems and intelligent digital assistants. It provides actionable guidance, practical checklists, and recommended practices to help builders embed safety, privacy, transparency, fairness, accessibility, and accountability throughout the AI lifecycle—from concept and design to deployment, monitoring, and continuous improvement.
>
> By adopting a Safety by Design approach, AI builders can proactively identify and mitigate risks, strengthen user trust, improve regulatory readiness, and develop AI solutions that are not only innovative and effective but also safe, inclusive, resilient, and aligned with Kenya's vision for responsible and human-centered AI.

Module 0

Module 0

Safety by Design: Guidance for AI Bot Builders

A THiNK AI Safety Resource for Innovators and AI/ML Developers

Introduction

AI bots are increasingly shaping how people access information, make decisions, and engage with services. With this GREAT power comes responsibility. Safety by Design (SBD) ensures bots are built with trust, responsibility, quality, inclusivity, and resilience at their core.

This guidance provides practical steps for bot builders to embed safety throughout the entire lifecycle of their projects—from concept to scale.

Why Safety by Design Matters

- Unsafe bots can spread misinformation, bias, or harmful content.
- Users may suffer emotional, social, or privacy harms.
- Reputational risks to developers and organizations are high.
- Safe bots build trust, adoption, and long-term impact.
- Safety is also a competitive differentiator in the AI marketplace.

Risks & Harms in AI Bots

a) Common Risks

- Misinformation or disinformation.
- Biased or discriminatory responses.
- Privacy breaches in conversations.
- Manipulation, prompt injection, or malicious misuse.

b) Potential Harms

- Emotional distress from offensive replies.
- Exclusion of marginalized users.
- Reputational harm to creators and organizations.
- Systemic harms at scale (e.g., misinformation campaigns).

c) Features That Strengthen Safety

• Guardrails in Dialogue – Refusal logic, content filters, and safe fallback answers.
• Bias Detection – Representative datasets and fairness audits.
• User Controls – Opt-out options, report abuse buttons, and transparency on the bot's scope.
• Privacy by Design – Minimal data retention, anonymization, and compliance with the Data Protection Act (DPA) and DKS 3007 AI Code of Practice.
• Accessibility & Inclusivity – Support for multiple languages, voice and screen readers, and plain language modes.
• Transparency – Clearly explain what the bot can and cannot do.

Embedding Safety Across the Bot Lifecycle

a) Design Phase

• Map risks and harms early.
• Involve diverse users in the design process.
• Define safety goals and standards.

b) Development Phase

• Train on diverse, safe datasets.
• Build safety guardrails into the code.
• Red-team bots with adversarial prompts.

c) Deployment Phase

• Conduct controlled rollouts (pilots).
• Monitor conversations in real time.
• Provide clear user onboarding with safety guidance.

d) Iteration Phase

• Perform continuous audits and updates.
• Collect and respond to user feedback.
• Align with the evolving AI Code of Practice.

Privacy-First Compliance

Consider the following insights when creating a compliance checklist.

a) Key Items to Consider

• Strong Foundation – PII detection, anonymization, and secure API architecture.
• Partial Compliance – Privacy disclaimers are present, but the UI banner and data retention policy are still missing.
• Gaps – Analytics continue to store direct user IDs, and retention scheduling has not been enforced.

b) Compliance Checklist Example

Area: Privacy Disclaimer
Requirement: Inform users that no PII is collected.
Status: Complete.
Notes: Implemented in the system prompt and footer; UI banner is still pending.

Area: Input Filtering
Requirement: Warn users about and filter PII.
Status: Complete.
Notes: Presidio-based detection with event sanitization.

Area: Metadata & Logs
Requirement: Minimize and anonymize metadata and logs.
Status: Partial.
Notes: User IDs are still stored directly; event sanitization has been improved.

Area: Retention
Requirement: Establish retention and cleanup policies.
Status: Not Complete.
Notes: Tooling exists, but no formal policy or automation has been implemented.

Area: Access Control
Requirement: Secure endpoints and logs.
Status: Complete.
Notes: Strong role-based access control (RBAC) and API key security.

Area: Audit Trail
Requirement: Document privacy compliance.
Status: Partial.
Notes: Logs exist, but a privacy-focused audit trail is still needed.

Area: Public Sector Ethics
Requirement: Align with the Data Protection Act (DPA) and accessibility requirements.
Status: Partial.
Notes: Awareness has been documented, but an accessibility audit is still missing.

Recommended Actions for Builders

• Add UI privacy disclaimers and a link to the privacy policy within the chatbot interface.
• Implement automatic log cleanup and retention policies (for example, retaining logs for 90 days).
• Anonymize user IDs used in analytics.
• Conduct formal accessibility audits in line with WCAG 2.1.
• Establish regular compliance reviews.

Validating Safety: The Conformity Assessment Process (CAP)

At Tech Innovators Network (THiNK), we use the CAP framework to validate safe bots.

1. Ownership Verification – Ensure responsible development.
2. Data Evaluation – Assess data quality, diversity, and bias mitigation.
3. Model Assessment – Test fairness, accuracy, and security.
4. Validation & Testing – Conduct compliance and safety checks.
5. Ongoing Monitoring – Maintain ethical and safe standards over time.

Proactive Risk Prevention

• Conduct red-team testing using harmful queries (e.g., self-harm and hate speech).
• Use human-in-the-loop escalation for sensitive topics.
• Perform regular safety updates and audits.
• Monitor misuse patterns while maintaining privacy safeguards.

Looking Ahead

• THiNK AI Safety Playbook (coming soon) – Practical tools for bot safety.
• NMWSO Safety by Design Curriculum (IREX Kenya) – Advanced training content.
• IREX chatbot on web.think.ke (currently undergoing testing) – Your go-to chatbot for Safety by Design queries. Feel free to engage with the chatbot.
• Ongoing resources and botathons for builders committed to responsible AI.

> ### “ Key Takeaway
>
> Safety by Design isn't a single feature—it's a mindset.
>
> By embedding safety at every stage, chatbot builders can engineer AI systems that are not only functional, but also safe, responsible, inclusive, and trusted by users.

Module 1

Module 1: Introduction to AI Safety

Module Overview

Artificial intelligence is transforming industries, improving service delivery, and creating new opportunities for innovation. However, alongside these benefits come significant risks that must be understood and managed. AI systems can unintentionally produce harmful outcomes through bias, misinformation, privacy violations, inaccurate predictions, or misuse by malicious actors. Understanding these risks is the first step toward building AI systems that are safe, responsible, and trustworthy.

This module introduces the fundamental concepts of AI safety and explores how safety differs from ethics and security. Learners will examine real-world examples from Kenya and around the world to understand the consequences of unsafe AI deployment and the importance of designing AI systems that protect people, respect human rights, and promote public trust.

What is AI Safety?

AI Safety refers to the practice of designing, developing, deploying, and maintaining artificial intelligence systems in ways that minimize risks to individuals, communities, and society while maximizing their benefits. Safe AI systems are reliable, transparent, ethical, fair, and accountable throughout their lifecycle.

AI safety encompasses several important dimensions:

Functional Safety focuses on preventing accidents, errors, and unintended consequences. For example, an AI system used in healthcare should provide accurate diagnoses and minimize the risk of harming patients through incorrect recommendations.

Ethical Safety ensures that AI systems operate fairly, respect human rights, avoid discrimination, and include diverse populations in their design and implementation.

Societal Safety addresses the broader impacts of AI on society, including misinformation, public trust, democratic processes, economic inequality, and long-term social risks.

Historical and Kenyan Case Studies of Unsafe AI Deployments

Examining real-world examples helps illustrate why AI safety is essential.

Kenyan Context

One of the most common examples involves **AI-powered credit scoring** used by digital lenders and fintech platforms. Some mobile lending applications rely on algorithms that may unfairly classify low-income earners, informal workers, or individuals with limited digital histories as high-risk borrowers. This can disproportionately affect women, youth, and rural communities by limiting their access to affordable financial services.

Another example is the growing use of **facial recognition technologies** in public spaces. AI-powered surveillance systems have been piloted in parts of Nairobi to improve public security and traffic management. While these technologies can support law enforcement, they also present risks such as privacy violations, inaccurate identification, false positives, and the potential targeting of marginalized communities if appropriate safeguards are not in place.

Kenya has also experienced the impact of **AI-assisted misinformation campaigns**, particularly during the 2017 and 2022 general elections, as well as during the 2024 and 2025 nationwide protests. AI-generated content, manipulated images, automated social media accounts, and deepfake technologies have contributed to the rapid spread of false information, undermining public trust and influencing public discourse.

Global Examples

Unsafe AI deployment has also been observed internationally.

In 2018, Amazon discontinued an AI recruitment system after discovering that it consistently favored male applicants because it had been trained using historical hiring data that reflected gender bias.

The COMPAS Risk Assessment System, widely used in parts of the United States criminal justice system, attracted criticism after studies found that its predictions disproportionately affected certain racial groups, raising concerns about fairness and accountability.

Microsoft's Tay chatbot, launched in 2016, was manipulated by users into generating offensive and harmful content within hours of its release, demonstrating the importance of robust content moderation and safety guardrails.

Tesla's Autopilot technology continues to highlight the risks associated with overreliance on autonomous systems, where incorrect identification of road conditions or driver overconfidence has contributed to several accidents.

Understanding the Difference Between Safety, Ethics, and Security

Although these concepts are closely related, they address different aspects of responsible AI.

AI Safety is concerned with preventing accidental harm and ensuring AI systems perform reliably. For example, an AI-powered health chatbot in Kenya should provide accurate medical information and avoid recommending harmful treatments based on poor-quality training data.

AI Ethics focuses on fairness, accountability, transparency, and respect for human rights. An example would be ensuring that AI-based loan approval systems do not unfairly disadvantage informal sector workers simply because they lack conventional financial records.

AI Security focuses on protecting AI systems from intentional attacks and malicious use. Examples include defending AI systems against hacking, preventing prompt injection attacks, and identifying AI-generated deepfakes that spread false political information during election campaigns.

A useful analogy is to compare these concepts with road transport:

- Safety is like wearing seatbelts and having airbags that protect passengers during accidents.
- Ethics is like ensuring that everyone has an equal opportunity to obtain a driver's licence through a fair process.
- Security is like protecting vehicles from theft, vandalism, or cyberattacks.

Together, safety, ethics, and security create trustworthy AI systems.

Key Risks in Artificial Intelligence

AI systems face several categories of risk that developers and organizations should understand.

Misuse occurs when AI technologies are intentionally used to spread misinformation, generate fake content, manipulate public opinion, or facilitate cybercrime.

Bias arises when AI systems produce unfair outcomes because of unrepresentative datasets or flawed design decisions. In Kenya, biased credit scoring and insurance algorithms may disadvantage people from low-income communities or those working in the informal sector.

Accidents happen when AI systems make unintended errors. Examples include diagnostic systems providing incorrect medical recommendations or automated systems making unsafe operational decisions.

Adversarial Manipulation involves deliberately deceiving AI systems. Examples include modified vehicle licence plates designed to confuse AI-powered traffic cameras or carefully crafted prompts intended to bypass chatbot safety controls.

Learning Outcomes

By the end of this module, learners should be able to:

- Explain the concept of AI safety and its importance in responsible AI development.
- Distinguish between AI safety, AI ethics, and AI security.
- Analyze Kenyan and international case studies involving unsafe AI deployments.
- Identify and categorize common AI risks and their potential impacts.
- Recommend practical strategies for reducing AI-related risks.

Learning Activities

1. Case Study Analysis

Learners will examine the use of AI-based credit scoring in Kenya's digital lending sector.

Working individually or in groups, learners should:

1. Review a summary describing how mobile lending platforms use AI to assess creditworthiness.
2. Identify the factors that contributed to unfair outcomes, such as biased datasets, limited transparency, or the over-penalization of informal workers.
3. Determine whether the issue primarily relates to AI safety, ethics, or security, while recognizing that multiple dimensions may overlap.
4. Recommend at least two practical measures that could reduce or prevent similar risks in future AI systems.

2. Discussion Questions

Learners should reflect on and discuss the following questions:

- Should fintech companies be required to demonstrate fairness before deploying AI-based credit scoring systems?
- Is algorithmic bias primarily a technical problem, or does it reflect broader inequalities within society?
- How can Kenya encourage AI innovation while ensuring that vulnerable populations are not excluded from essential services?

Teaching Materials

The instructor should provide presentation slides covering key AI safety concepts, comparisons between safety, ethics, and security, AI risk categories, and both Kenyan and international case studies.

Recommended readings include *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* by Brundage et al. (2018), together with articles, policy papers, or reports discussing AI bias and access to financial services in Kenya.

Suggested multimedia resources include news reports on AI-powered surveillance technologies in Nairobi and documentaries or news segments examining misinformation during Kenyan elections.

> “
>
> # Assessment
>
> Learners' understanding will be assessed through multiple approaches.
>
> A short quiz will test their ability to distinguish between AI safety, ethics, and security using practical Kenyan examples.
>
> Learners will also complete a reflection essay discussing the AI risk they believe poses the greatest challenge to Kenya and explaining their reasoning.
>
> Participation during case study discussions will contribute to the final assessment, with emphasis placed on critical thinking, evidence-based reasoning, and the ability to propose practical solutions for safer AI systems.