WHITE PAPER

Safety as Innovation

An African Stack for Trusted AI

AICE Africa and the Eswatini Standards Authority (ESWASA)

November 2026

www.aiceafrica.com | www.eswasa.co.sz

About this paper

How to cite this paper: AICE Africa and Eswatini Standards Authority (2026). *Safety as Innovation: An African Stack for Trusted AI*. White paper. Mbabane and Nairobi: ESWASA and AICE Africa.

About this paper: This white paper sets out a framework and a set of proposals for discussion with governments, regulators, standards bodies, industry, and development partners. It is not a legal opinion or a technical standard. Programme dates after February 2027 are targets and depend on funding, accreditation, and partner decisions. Monetary figures are in US dollars unless stated otherwise. Sources were accessed in November 2026. Disclosure: AICE Africa provides AI products, training, and advisory services; the impartiality arrangements set out in Section 9 govern its role in the Centre and the proposed lab.

Version 1.0, November 2026. Released in Vienna at the AIM Global Action Forum 2026 (25–26 November 2026), held as part of UNIDO's 60th anniversary. Updates and contact: www.aiceafrica.com and www.eswasa.co.sz.

© 2026 AICE Africa and Eswatini Standards Authority (ESWASA).

Contents

Foreword 2
Executive summary 3
1. Why now: AI as an equaliser, and why safety decides it 5
2. The AI safety stack 6
3. Safety as innovation, not a list of dos and don'ts 9
4. Global voices, and where Africa stands 11
5. Layers of responsibility 12
6. New roles every organisation needs 14
7. Independent AI assurance: the sector's key value 15
8. Sector safety profiles 17
9. Eswatini: the test bed the world needs 19
10. Aligning with global standards, and trading AI across borders 22
11. Capacity development 25
12. The future of AI safety: a decentralised, more capable optimiser 26
13. The opportunity stack for Africa 29
14. Roadmap and calls to action 29
Acronyms 32
About the authors 34
Sources 35

Foreword

Africa does not have to choose between using AI and being safe with it. Our view is that safety is how we use it fully, and that the countries that can test, certify, and trust AI systems will be the ones that capture their value.

In October 2026, the Eswatini Standards Authority (ESWASA) and AICE Africa launched the AI Safety and Standards Centre of Excellence in Eswatini. We did it because we saw a gap. AI is arriving in African banks, hospitals, factories, schools, and ministries faster than African institutions can evaluate it. Most of the systems we depend on are built, tested, and certified elsewhere. That is a sovereignty problem, a consumer protection problem, and, we believe, a missed economic opportunity.

This paper sets out how we see AI safety: as a stack of technical and institutional layers, as a shared responsibility from government to the individual user, and above all as an innovation and a market in its own right. It draws on our work in Eswatini, on six years of AICE Africa's work training engineers and leaders and deploying AI with enterprises and governments across the continent, and on the global debate, including the warnings of researchers such as Nate Soares.

We offer this paper as an invitation to governments and regulators, standards bodies across ARSO, enterprises and model developers, our partners in Europe, the United States, Asia, and beyond, and the African innovators who will build the safety industry itself. We are proud to present it in Vienna at UNIDO's AIM Global Action Forum 2026, as UNIDO marks its 60th anniversary, and to have UNIDO with us as a partner in this work.

John Kamara, Chief Executive Officer, AICE Africa

Ncamiso Mhlanga, Executive Director, Eswatini Standards Authority (ESWASA)

Executive summary

AI safety should be treated as an engine of innovation and trade, not a list of dos and don'ts. Safety that can be measured, tested, and certified lets buyers, regulators, and citizens say yes to AI with confidence. Africa can build that capability now, with Eswatini as the first test bed.

Key messages

- **Safety is a market, not only a cost.** As with cybersecurity, AI assurance creates professions, firms, and exports, and Africa can be a producer of it.
- **Seven layers, one stack.** Cybersecurity and the data and model supply chain are the foundation; responsible AI and ownership is the governance around everything.
- **Everyone owns a part.** From the AU and regulators to model developers, deployers, and citizens, responsibility follows control, and every organisation needs named AI safety roles.
- **Test it, certify it, trust it, trade it.** Accredited testing and certification, recognised across borders, turn safety into market access under the AfCFTA and beyond.
- **Eswatini is the test bed.** A national Centre of Excellence, an independent testing lab, a Policy Studio, and two regional gates in 2027 carry the work from one country to the continent.

The problem. AI adoption is outrunning the safeguards around it. The AI Incident Database recorded 362 documented incidents in 2025, up from 233 in 2024 (Stanford AI Index 2026). Agents now act, spend, and decide on people's behalf. Africa is not starting from zero: Kenya was a founding member of the International Network of AI Safety Institutes in 2024, and research hubs such as the African Hub on AI Safety, Peace and Security at the University of Cape Town are active. But dedicated national testing capacity remains scarce (Toward an African Agenda for AI Safety, 2025), and most systems used on the continent are evaluated abroad, if at all.

Our thesis. Safety is what enables AI's value to be used. Cybersecurity shows the pattern: as digital risk grew, an entire assurance industry grew alongside it, and trust became something people pay for. AI assurance is following the same path, and Africa can be a producer rather than only a consumer. Our shorthand is: **test it, certify it, trust it, trade it.**

The stack. We see seven complementary layers of AI safety: cybersecurity; alignment; agency and Know Your Agent; responsible AI and ownership; self-improving systems; data and model supply chain; and societal resilience. Cybersecurity and the supply chain form the technical foundation, and responsible AI and ownership is the governance layer around all the others. Each layer has its own science, its own controls, and its own market.

Shared responsibility. Safety runs from governments and regulators, through standards bodies, model developers, platforms, enterprises, and deployers, to the individual user. Every organisation using AI now needs named people: at minimum three core AI safety roles (an AI Safety and Standards Officer, an AI

Ethics and Compliance Officer, and an AI Monitoring and Technical Expert), within the roles set out in Section 6. Independent assurance will become a sector of its own.

Sector by sector. Healthcare, industrial AI and energy, financial services and mobile money, government, education, transport and logistics, agriculture, and telecoms and media each carry different risks. Each sector should define its own safety profile with its regulators and industry leaders, on a common base.

Eswatini will serve as the test bed. ESWASA and AICE Africa are designing and seeking funding for an independent AI testing lab, building a Policy Studio that maps African, European, and global rules onto technical tests and evidence, and training regulators, boards, and practitioners. Pilot evaluations of a small number of Eswatini systems will run from early 2027 to mid-2027, before the lab opens. The lab targets initial operations in late 2027 and accredited operations in 2028. The 1st SADC Convening on AI Safety, Standards, and Ethics (23–24 February 2027) and the SHIELD 2027 Africa AI Safety and Standards Summit (28–29 July 2027) will carry this from national to regional scale. At SHIELD 2027, we will table a zero-draft Pan-African AI Safety and Standards Charter, developed for discussion with ARSO.

The prize. Test once, trust everywhere, trade across borders. If test results from an accredited lab in Mbabane are recognised in Nairobi and Lagos and accepted by buyers and regulators in Brussels as credible evidence, African safety becomes exportable and African AI becomes tradable. That is how AI becomes an equaliser for us.

What we ask: Governments to plan for AI safety capacity in every ministry and regulator. Enterprises to appoint AI safety roles and test before deployment. Accreditation bodies to extend their recognised scopes to AI. Global partners to recognise African testing and co-build standards. Funders, including the AfDB and UNDP through the AI 10 Billion Initiative, to back the lab, training, and regional convening, and to treat assurance as part of AI readiness. Open-source foundations and communities, such as the Linux Foundation, and industrial partners, such as UNIDO, to build safety into the tools and factories Africa will use.

1. Why now: AI as an equaliser, and why safety decides it

AI can be the first general-purpose technology that reaches Africa at the same moment it reaches everyone else. Whether it narrows or widens the gap depends on who can trust, test, and own the systems.

The upside is large. Top-down scenarios put AI's contribution to African economies at $1.2–1.5 trillion by 2030. SAP's $1.5 trillion figure assumes Africa captures 10% of the global AI market (SAP); Microsoft's $1.2 trillion is Africa's share of a $15.7 trillion global projection (Microsoft via AllAfrica). By contrast, McKinsey's bottom-up sectoral estimates see up to $100 billion per year from generative AI alone if institutions address barriers and build for scale (McKinsey). The two methods answer different questions. Africa accounts for about 3% of the world economy today (ISS African Futures), so the top-down figures are scenarios of what is possible, not baselines. Each depends on adoption that people and regulators trust.

Africa's own development institutions are calling for an AI-driven Africa and putting money behind it. In February 2026, the African Development Bank and UNDP launched the AI 10 Billion Initiative at the Nairobi AI Forum (AfDB via AllAfrica). By October 2026, it had moved from launch to delivery, aiming to orchestrate about $1 billion a year over the next decade and contribute to 25 million new jobs by 2035. More conservatively, AfDB research projects that inclusive AI adoption could add up to $1 trillion to Africa's GDP by 2035 (TechAfrica News). The G7-endorsed AI Hub for Sustainable Development, run by UNDP with Italy's support, is tackling the compute gap; at the same forum, 1.5 million GPU hours were announced for 130 African innovators (UNDP). Capital at this scale requires confidence that the systems it finances are safe to deploy. Safety institutions are what make AI investment bankable.

The risk is just as concrete. The International AI Safety Report 2026, written by more than 100 experts and chaired by Yoshua Bengio, finds criminals using general-purpose AI in cyberattacks, deepfakes used for fraud and scams, and agents becoming a major focus of development. Africa is home to 18% of the world's population but less than 1% of global data-centre capacity (World Economic Forum), so most models serving Africans are trained and hosted elsewhere, on data that rarely reflects African languages and lives.

Three reasons make this the moment to act:

1. **We are early.** Most African institutions are still choosing their first AI systems. Building safety in now costs far less than retrofitting it after failures, as the cyber industry learned.
2. **Agents change the risk.** Systems that answer questions are one thing. Systems that move money, approve claims, or route ambulances are another. The question shifts from 'Is the output right?' to 'Who authorised this agent, and can we stop it?'
3. **Rules are being written now.** The EU AI Act's high-risk obligations apply from 2 December 2027 for stand-alone systems and 2 August 2028 for AI in regulated products, under amendments in force since July 2026 (Regulation (EU) 2026/1744). The first harmonised standard, EN 18286:2026 on quality management, was cited in the Official Journal in October 2026 (Implementing Decision (EU) 2026/2274); most of the others are still being drafted. If Africa is not in the room, it will be governed by rules it did not shape.

An equaliser needs three things: access to the technology, the ability to judge it, and the right to own what it produces. Safety institutions supply the second and protect the third.

2. The AI safety stack

AI safety is not one problem but seven related ones. We treat them as complementary layers rather than a strict hierarchy. Cybersecurity and the data and model supply chain are the technical foundation. Alignment, agency, and self-improvement govern how systems behave. Societal resilience covers harms that only appear at scale. Responsible AI and ownership is the governance layer around all of them. Each layer can fail on its own, each needs its own tests, and each is a place where products and jobs will be created.

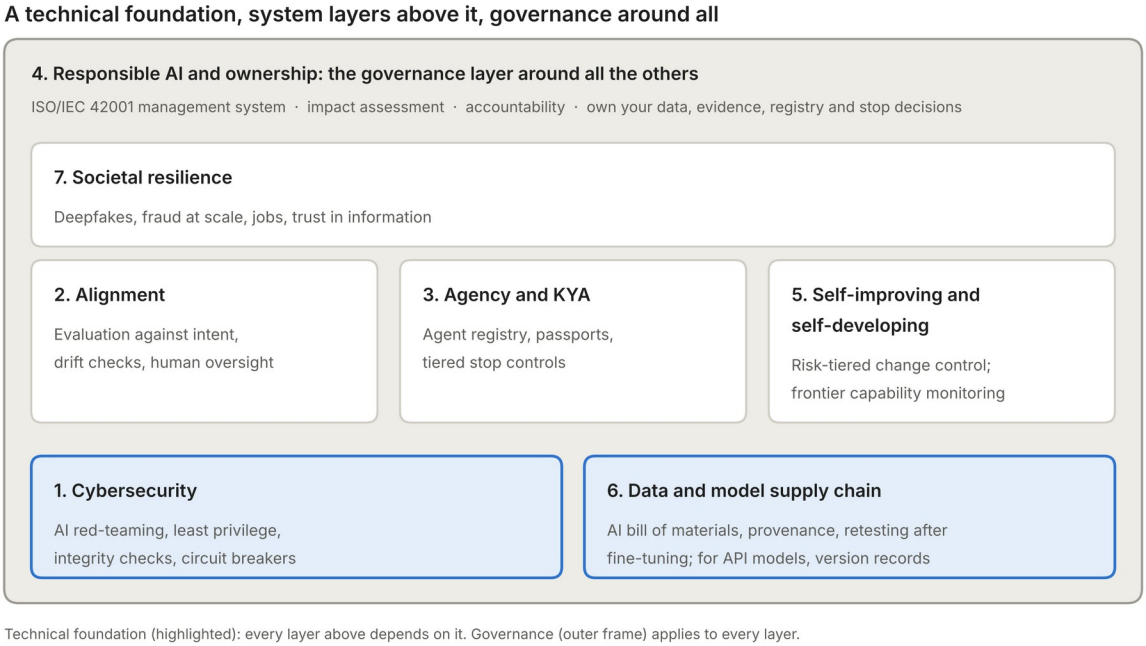


Figure 1. The AI safety stack. Seven complementary layers: a technical foundation, system layers above it, and governance around all. Layers are numbered in the order discussed below.

Layer 1: Cybersecurity

An AI system that can be breached cannot be safe, no matter how well aligned. AI adds new attack surfaces to existing ones: prompt injection, data poisoning, model theft, jailbroken guardrails, and agents tricked into using their tools against their owners. It also arms attackers: in 2025, an AI agent placed in the top 5% of teams in a major cybersecurity competition (International AI Safety Report 2026, Office of the Chair). Public references provide a shared vocabulary for this layer: the OWASP Top 10 for LLM and generative AI applications (OWASP GenAI); MITRE ATLAS, a catalogue of adversary tactics across the machine-learning life cycle; and ISO/IEC 27090:2026, international guidance on security threats to AI systems.

What safety looks like: AI-specific red-teaming, structured against MITRE ATLAS and OWASP, alongside the annual pen test; least-privilege access for every model and agent; cryptographic integrity checks against tampering and poisoning of models and data (provenance and vendor governance sit in Layer 6); AI-driven monitoring that matches AI-driven attacks, backed by independent, non-AI circuit breakers such as hard rate and spending limits.

Layer 2: Alignment

Alignment asks whether a system pursues the goals we meant, not just the ones we measured. Modern models are, in Nate Soares's phrase, grown rather than crafted, so their behaviour has to be tested rather than assumed. Day to day, misalignment looks like Goodhart's law: a collections agent that hits its recovery target by harassing customers, or a triage model that optimises waiting times by under-referring.

What safety looks like: evaluations against intended outcomes, not just metrics; behavioural fingerprints and drift detection after deployment; human oversight where stakes are high; interpretability research where it is mature enough to use.

Layer 3: Agency and Know Your Agent

Once AI acts, the central questions become identity and authority. Know Your Agent (KYA) applies Know Your Customer (KYC) logic to software: which agent is acting, who authorised it, and within what limits (PYMNTS, 2026a). Visa, Mastercard, and Ant International are now developing an interoperable KYA framework for agentic commerce (PYMNTS, 2026b). KYA is an industry practice today, not yet a legal requirement.

The protocols agents use to reach tools, data, and each other are converging on open, neutrally governed standards. In December 2025, the Linux Foundation formed the Agentic AI Foundation as a neutral home for Anthropic's Model Context Protocol (MCP), Block's goose, and OpenAI's AGENTS.md (Linux Foundation, 2025); Google's Agent2Agent (A2A) protocol is also hosted at the Linux Foundation (InfoQ). Open protocols are the natural place to build in KYA checks, so that identity and permissions travel with every agent call.

What safety looks like: a registry of every agent an organisation runs; a passport for each agent stating its owner, permissions, tools, and spending limits, with limits that also apply in aggregate across any sub-agents it delegates to; tiered stop controls, tested regularly: local runtime shutdown by the deploying organisation, rapid revocation of an agent's credentials at payment rails and API gateways so that counterparties refuse it, and graceful fail-safe degradation for multi-agent workflows and safety-critical systems. Logs should let any incident be replayed. Stopping an agent does not undo what it has already done: uncommitted actions can be held or cancelled, but a settled payment or a binding instruction needs a refund, compensating action, or dispute process, which is why the strongest controls act before money moves.

Layer 4: Responsible AI and ownership

Responsible AI covers fairness, transparency, privacy, explainability, and accountability. It is also the governance layer that wraps all the others: an AI management system aligned with ISO/IEC 42001 sets the

policies, risk assessments, and change control that apply across every technical layer. We add ownership because responsibility without control is a slogan. An organisation should retain authority over four things: its data, its evidence, its agent registry, and the decision to stop its own systems. For Africa, ownership also means African languages, African data, and African IP inside the systems we use.

What safety looks like: an AI management system aligned with ISO/IEC 42001; impact assessments before deployment; clear accountability for each system; contracts that keep data and evidence under local control.

Layer 5: Self-improving and self-developing systems

AI is increasingly used to write code, design experiments, and train other AI. Systems that improve themselves, or are improved by other AI, can change faster than humans can review them. This is the layer where Soares and many frontier labs see the gravest long-term risk, and where today's tools are weakest.

What safety looks like: for enterprise deployers, sandboxing autonomous coding and workflow agents, prohibiting unreviewed self-modification or self-fine-tuning in production, and applying risk-tiered regression testing when model versions, tools, or permissions change, with a full re-evaluation reserved for material changes. For frontier labs and regulators, compute and capability monitoring, verified capability disclosures from model providers, and participation in international work on frontier risk thresholds.

Layer 6: Data and model supply chain

Most organisations do not build models; they assemble them from foundation models, open-weight models, plug-ins, APIs, and data vendors. Each component is a dependency, and an open-weight model can be fine-tuned to remove its safeguards. This layer sits alongside cybersecurity at the foundation: if weights or data are compromised, the alignment and agency controls built on top of them cannot be trusted. Deployers also depend on publishers for information they cannot produce themselves, such as pre-training data summaries and base-model evaluations.

Open-source tooling is making this practical. The Linux Foundation's Open Source Security Foundation (OpenSSF) released version 1.0 of its model-signing tool in April 2025 and formalised it in June 2025 as the OpenSSF Model Signing (OMS) specification, a standard for cryptographically signing and verifying machine-learning models (OpenSSF, 2025). In September 2026, the Open Secure AI Alliance moved to the Linux Foundation to build a shared, open defence stack for AI (Linux Foundation, 2026), with contributions such as zero-trust identity frameworks, secure model formats, and model scanning and auditing tools (ARC Advisory Group). African labs and startups can adopt these tools rather than rebuild them and contribute African test packs in return. A signature proves that a model file has not changed; it says nothing about the model's behaviour, so signing complements behavioural testing rather than replacing it.

Many African organisations will never hold the weights at all. A bank that uses a proprietary model through a foreign cloud API cannot inspect or sign it and may not know when the provider changes it. For these deployments, assurance rests on records of the model version and configuration in use, contractual transparency and change notification from the provider, access controls and logging at the organisation's own gateway, independent behavioural testing of the system as deployed, and drift monitoring in

production. Where a provider cannot guarantee a fixed version, the deployer should record that and retest more often.

What safety looks like: an AI bill of materials for every system, recording base-model, dataset, and adapter hashes for open weights, or provider, version, and configuration for API models; provenance checks on data and weights; behavioural regression testing after any fine-tuning or model change; vendor assurance clauses; and upstream disclosures from model publishers.

Layer 7: Societal resilience

Some harms only appear at scale: deepfake fraud, election manipulation, mass job change, concentration of power, and loss of trust in what we see and hear. No single company can test for these.

What safety looks like: national incident reporting and sharing, with a confidential, non-punitive channel for near misses on the aviation and CERT model, so that failures are reported rather than hidden; content provenance and detection; AI literacy for the public; observatories that track downstream impact across a sector or country.

3. Safety as innovation, not a list of dos and don'ts

The most useful way to think about AI safety is as the brakes on a fast car: they are what let you drive fast. Rules that only say 'don't' slow adoption. Safety that can be measured, tested, and certified speeds it up because it lets buyers, regulators, and citizens say yes with confidence.

The cybersecurity benchmark

Cybersecurity is the clearest example. A decade ago, it was a cost centre within the IT department. Today, it is a board-level discipline, a regulated function, and one of the largest technology markets in the world.

Table 1. Cybersecurity as a precedent for AI assurance

Measure	Around 2015	Around 2025
Global cybersecurity market	About $75 billion (CSO)	About $228 billion (MarketsandMarkets)
Estimated global cost of cybercrime	$3 trillion	$10.5 trillion (Cybersecurity Ventures)

Figures come from different analysts and are indicative. Cybercrime costs are estimates and projections.

As the threat grew roughly threefold, so did the defence industry. Digital risk had already given rise to new professions (CISOs, SOC analysts, pen testers), standards (ISO/IEC 27001), institutions (CERTs, cyber agencies), and companies; the past decade turned them into a global industry.

AI assurance is following the same path, earlier in its curve. The UK's AI assurance market already includes an estimated 524 firms, employs more than 12,500 people, and contributes about £1.01 billion annually, and could reach about £6.5 billion by 2035 if recent growth continues. The UK's mature cyber assurance

ecosystem is worth nearly £4 billion (UK DSIT, 2024). BSI alone issued 11 ISO/IEC 42001 certificates in 2025 and reached its 50th in October 2026, less than two years after its first (BSI, 2026).

Where the economic value of safety comes from

1. **Faster adoption.** A tested and certified system clears procurement, regulators, and boards more quickly. Trust shortens sales cycles.
2. **Avoided losses.** Fraud, outages, recalls, fines, and lawsuits cost more than prevention. Under the EU AI Act, fines can reach up to €15 million or 3% of worldwide turnover for breaches of high-risk obligations, and up to €35 million or 7% for prohibited practices (EU AI Act, Article 99). The Act applies to African organisations within its scope, for example, by placing systems on the EU market or having their outputs used in the EU.
3. **Market access.** Evidence of compliance is a passport. A system tested to recognised standards can be sold across borders.
4. **New services.** Testing, red-teaming, monitoring, certification, insurance, training, and incident response are businesses in their own right.
5. **Better products.** Testing finds failure modes early. Safe AI is usually also more reliable AI.

Connecting policy to the opportunity

Why assurance markets form. The economics is old and well-tested. When buyers cannot distinguish a safe product from an unsafe one, they discount every product, driving good sellers out: the problem George Akerlof described in 'The Market for "Lemons"' (Akerlof, 1970). Independent testing and certification solve this by making quality visible. Because they lower the cost of every later transaction, assurance grows in step with the market it serves. For AI, where even builders cannot fully explain their systems, the information gap is wider than in most markets, and so is the value of closing it.

Policy creates demand; capability creates supply. When a regulator requires evidence, someone must be able to produce it. Our approach is to design the two together: every rule we help write should include the test that proves compliance, the people trained to run it, and the local firms able to sell it. That turns regulation from a cost imposed on business into a market that business can serve.

We use a simple shorthand for this: **test it, certify it, trust it, trade it.** Each step has a different owner and a different output (Figure 2).

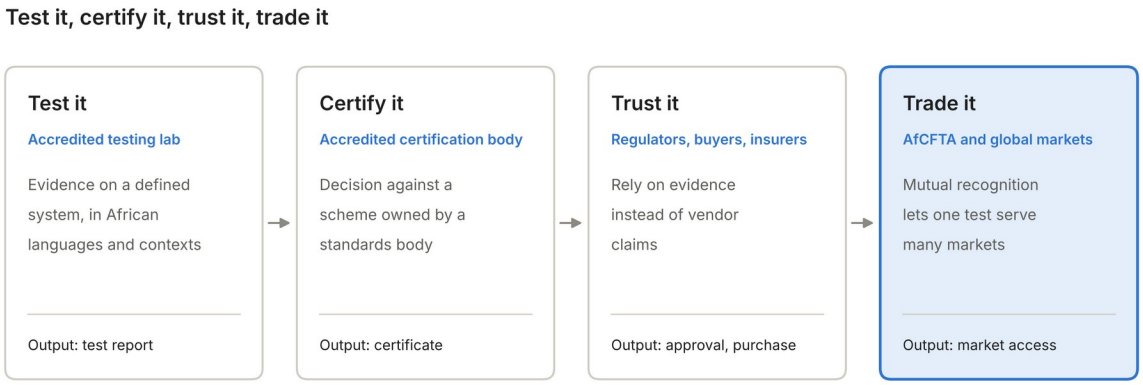


Figure 2. From test to trade. Each step has its own owner and output; mutual recognition lets one accredited test serve many markets.

4. Global voices, and where Africa stands

The world's leading safety thinkers disagree on how much danger lies ahead, but they agree on one thing: we cannot trust what we cannot test. That is the common ground Africa can build on.

Nate Soares: the strongest warning

Nate Soares, president of the Machine Intelligence Research Institute (MIRI), co-wrote *If Anyone Builds It, Everyone Dies* with Eliezer Yudkowsky (Yudkowsky and Soares, 2025; Fortune). They argue that today's AI systems are grown rather than crafted, so we cannot see inside them or reliably set their goals (Future of Life Institute). They conclude that if anyone builds a superintelligence with current methods, the most likely outcome is catastrophe. The title begins with 'if'; it is a conditional warning, not a prophecy.

Soares does not call for an end to all AI. Speaking at Harvard in March 2026, he drew the line at 'superintelligence that no one understands' (The Tech). MIRI's technical governance team has proposed an international agreement to halt progress towards superintelligence while preserving access to current, beneficial AI applications (MIRI). Its model borrows from nuclear non-proliferation: compute thresholds, monitoring, and verification.

The broader scientific picture

The International AI Safety Report 2026, chaired by Yoshua Bengio and guided by an expert panel nominated by 29 countries, the UN, the OECD, and the EU, does not make policy recommendations. It finds rapid capability gains, the active criminal use of AI, and an ongoing challenge in building effective safeguards. Many frontier developers now publish their own safety frameworks with capability thresholds and 'if-then' commitments. Critics of the Soares and Yudkowsky thesis, such as Adam Becker in *The Atlantic*, argue that the thesis makes unwarranted logical leaps and that present harms deserve more attention than speculative ones.

Our position

We take the frontier warnings seriously, and we see no conflict between them and safety as innovation. Our reasoning:

- **Any global agreement needs downstream as well as upstream verifiers.** Upstream verification of frontier training (chip tracking, data-centre monitoring, compute thresholds) will happen where hyperscale clusters are, and Africa hosts few of them. Downstream deployment assurance is where African regulators have direct jurisdiction, and African labs can build it now: verifying the provenance of models submitted for testing, confirming that regulated deployers have not stripped safeguards from models they fine-tune, testing closed models reached through APIs, tracking post-deployment incidents, and supplying trained evaluators to multilateral inspection rosters. Detecting covert fine-tuning on hardware no one inspects, or monitoring compute in foreign clouds, will need research and international agreements that do not yet exist; African labs can prepare for those roles but should not claim them yet.
- **Today's harms are real and local.** A biased credit model, a hallucinating clinical assistant, or a hijacked payments agent harms African citizens now. These need testing now.
- **Safety must be a science.** Soares's core point is that we lack the understanding needed to make systems safe. The answer is more measurement, more evaluation, and more research, which is exactly what labs and assurance markets fund.
- **Africa should have a voice.** As a group of countries that do not build frontier models, Africa has a shared interest in rules that keep the most dangerous development in check while keeping beneficial AI open to all. African institutions should be part of that conversation, not observers.

In short: we agree that some lines should not be crossed without understanding. We believe the way to know where the lines are is to build the science and institutions to test for them, and that Africa can help build both.

5. Layers of responsibility

No single actor can make AI safe. Responsibility must be shared so that each actor owns what it can actually control, and gaps between them are closed rather than debated.

Table 2. Who controls what, and who is responsible for what

Actor	What it controls	What it is responsible for
Continental and regional bodies (AU, RECs such as SADC, ARSO, AFRAC)	Common frameworks, harmonised standards, mutual recognition	A Pan-African AI Safety and Standards Charter; one test, many markets; a shared African voice in global forums
National governments	Law, strategy, public procurement, public services	National AI strategy with a safety pillar; AI safety capacity in every ministry; safe AI in its own services; incident reporting
Regulators and standards bodies (central banks, health, telecoms, data protection authorities, ESWASA, and peers)	Sector rules, standards, certification schemes, market access	Development and adoption of AI standards; sector safety profiles; certification schemes aligned to ISO/IEC 42001; sandboxes; supervision of AI in their sectors
Accreditation bodies (such as SADCAS, under AFRAC and Global Accreditation Cooperation Incorporated)	Accreditation of labs and certifiers	Assessing the competence and impartiality of AI testing labs and certification bodies, and extending recognised scopes to AI, so results can be recognised across borders
Frontier and foundation model developers	Training, model weights, model cards, usage policies	Pre-release evaluation; disclosure of capabilities and limits; frontier safety frameworks; support for independent testing, including in African languages
Open-weight model publishers	Pre-training data, base weights, release decisions, licences	Pre-release risk assessment; upstream disclosures (pre-training data summaries, base-model evaluation cards, bills of materials) that downstream deployers need for their own checks; release practices that weigh misuse against openness
Proprietary AI platforms and cloud providers	Access, APIs, hosting, logs, model versions	Abuse monitoring; data residency options; notice of model changes to deployers; audit access for regulators and accredited labs
AI organisations, integrators, and startups	Products built on models	Secure design; testing before shipping; AI bill of materials; honest claims
Enterprises and public bodies deploying AI	Use cases, data, agents, people	Named AI safety roles; agent registry; impact assessment; monitoring; tested stop controls; staff training
Organisations running open-weight models themselves	Fine-tuning data, system prompts, deployment, guardrails, and hosting (not the pre-training corpus or base weights)	All deployer duties, plus retesting after every fine-tune and securing weights against theft or tampering
Independent assurance providers	Tests, audits, certificates	Independence, competence, and accredited methods; no testing or certification of clients they have advised or systems they have configured
Individuals and civil society	Their own use, their voice	AI literacy; reporting harms; scrutiny of public and private AI

Two points deserve emphasis.

First, **deploying an open-weight model shifts operational responsibility to the deployer, but not the duty of upstream transparency. The EU AI Act (Regulation (EU) 2024/1689) keeps two regimes separate: one for general-purpose models and one for high-risk systems**. A bank that fine-tunes an open-weight model for a high-risk purpose, such as creditworthiness assessment, becomes, under Article 25, the provider of that high-risk AI *system*, not of the base model. It is responsible for post-modification safety and must retest after fine-tuning. Yet it cannot see the pre-training corpus or detect every latent flaw in the base weights, and Article 53(2) exempts free and open-source general-purpose models without systemic risk from providing technical documentation to downstream integrators. Our call for publishers to share evaluation cards and bills of materials is therefore a deliberate proposal that goes beyond the EU baseline, enabling deployers to complete their supply-chain checks.

Second, **responsibility follows control, but accountability cannot be outsourced.** An enterprise can buy a system that has been independently tested, but it still owns the decision to use it, the data it feeds in, and the harm if it fails.

6. New roles every organisation needs

Every organisation that uses AI should now plan for named AI safety roles, just as every bank came to need a CISO. The trend is already evident: AI-specific governance roles grew by 17% in 2025, according to McKinsey survey data reported in the Stanford AI Index 2026.

Table 3. Six AI safety roles and the lines of defence

Role	Line	Core job	Reports to
AI Safety and Standards Officer	Second line (oversight)	Owns the AI safety framework, risk thresholds, and independent oversight of the agent registry; can require a system to be stopped when thresholds are breached; leads ISO/IEC 42001 readiness and the link to regulators	CEO or board risk committee
AI Ethics and Compliance Officer	Second line (oversight)	Impact assessments, fairness, privacy, regulatory mapping across jurisdictions, and complaints and redress	General counsel or chief compliance officer
AI Monitoring and Technical Expert	First line (operations)	Live monitoring, drift detection, AI cybersecurity, incident response, and operation of stop controls	CISO or CTO, with an escalation line to the AI Safety and Standards Officer
AI Agentic Lead	First line (build)	Designs and configures agents, their permissions, and their passports, within thresholds set by the AI Safety and Standards Officer	CTO or COO

Role	Line	Core job	Reports to
AI Architect	First line (build)	Ensures systems are built safely by design: data flows, model choice, supply chain	CTO
AI Red Team (People and Agents)	Independent challenge	Attacks the organisation's own AI before others do	AI Safety and Standards Officer, with a line to the board risk committee

The first line builds and runs AI; the second line sets limits and independently checks. The red team sits outside the team that operates the defences it tests, so it does not mark its own work. Above them, internal audit, as the third line, provides the board with independent assurance that the entire AI management system works, including the second line. This is the organisation's existing audit function, extended to AI, not a new department. Organisations without an internal audit function can use an independent external assessor for the internal audits required by ISO/IEC 42001.

Smaller organisations will not hire six people. They can combine roles, share them across a group, or buy specialist support as a service. Four rules keep this safe:

1. **Combine like with like.** Merge build roles with build roles and oversight roles with oversight roles, so that no one grades their own system. The monitoring role should not be merged with the people who build the agents it watches.
2. **Keep the decisions inside.** Monitoring and advice can be bought, but final sign-off on agent permissions and the authority to stop a system must stay with a named internal executive.
3. **Keep advisers and certifiers separate.** A provider that supplies outsourced safety governance to an organisation cannot also serve as that organisation's independent certifier.
4. **Never be the sole judge of your own controls.** Where one person must both operate controls and assess them, a documented independent review must compensate.

Boards need literacy, too. A board that cannot ask, 'Which agents act on our behalf, and who can stop them?' cannot oversee AI risk. AICE Africa's enterprise AI workshops for CEOs, CTOs, and boards, already under way with banks and corporates in Kenya, Rwanda, and across the region, are designed to close this gap.

7. Independent AI assurance: the sector's key value

The organisation that builds or sells an AI system should not be the only one to say it is safe. Independent assurance by companies and labs with no stake in the result turns a vendor's claim into evidence that a buyer or regulator can rely on.

We expect an African AI assurance industry to form around eight service lines:

1. **Pre-deployment testing and red-teaming**, including in African languages and contexts, building on African benchmarks (Section 8) and structured against MITRE ATLAS and the OWASP GenAI Top 10.

2. **Certification**, issued by accredited certification bodies on two tracks: organisational AI management systems against ISO/IEC 42001, and, over time, individual high-risk systems and products, using test reports from accredited labs as evidence.
3. **Continuous monitoring** of live systems for drift, misuse, and attack as an assurance service distinct from accredited testing.
4. **Agent identity and KYA services**: registries, passports, and verification.
5. **Incident response and forensics**: containing a failure, replaying it, and finding its cause.
6. **Regulatory mapping and compliance**: translating the EU AI Act, national laws, and sector rules into controls.
7. **AI risk insurance and its underwriting data.**
8. **Training and professional certification** of assessors, auditors, and AI safety officers through accredited schemes for certifying personnel.

The UK found that 80% of its specialised AI assurance firms showed signs of growth, and it is now supporting the development of an AI assurance profession (UK DSIT, 2025). Africa can develop the same profession, with three advantages: lower costs, deep knowledge of local languages and conditions, and the chance to design for agents and open-weight models from day one rather than retrofitting. No comparable estimate of African demand for AI assurance exists yet; the pilot evaluations and the Centre's work will help build that evidence.

For this market to be trusted, it needs the same conformity-assessment infrastructure that underpins trade in goods:

Table 4. Conformity-assessment infrastructure for AI

Activity	What is assessed	Governing standards	Who accredits
Management-system certification	An organisation's AI governance, risk, and change processes	ISO/IEC 42001; ISO/IEC 42006 and ISO/IEC 17021-1 for certification bodies	National or regional accreditation body
Technical testing	A specific model, agent, or system: robustness, security, fairness, classification performance, language performance	ISO/IEC 17025 for testing labs, with methods drawn from published standards such as ISO/IEC TR 29119-11 (testing AI systems), the ISO/IEC 24029 series (neural-network robustness), ISO/IEC TR 24027 (bias) and ISO/IEC TS 4213 (classification performance)	SADCAS in SADC, recognised under AFRAC and Global Accreditation Cooperation Incorporated for its current scopes
Product or system certification	Whether a specific system meets the requirements of a defined certification scheme	ISO/IEC 17065 for certification bodies; ISO/IEC 17067:2026 for designing the scheme	National or regional accreditation body

Accreditation turns a local certificate into one that other countries accept, and it operates scope by scope. Since January 2026, the two global accreditation networks, IAF and ILAC, have operated as a single body, Global Accreditation Cooperation Incorporated, or Global ACI (Global ACI, 2026). SADCAS, the multi-economy accreditation body for SADC, holds international recognition for ISO/IEC 17025 testing (SADC; SADCAS, 2019b), for several management-system certification scopes (SADCAS, 2019a), and, since December 2023, for product certification (IAF, 2024). This makes it the natural route for AI testing and certification in the region, but none of its recognised scopes yet cover AI; extending them is part of the work ahead.

Product and system certification also needs a scheme with defined requirements, thresholds, and decision rules, owned by a scheme owner, such as a national standards body. The lab produces evidence; a separately accredited certification body makes the decision. Even then, accreditation does not automatically confer acceptance: each regulator decides which results it will rely on. The market also needs strict separation between those who advise and those who test or certify, as well as public registers of certificates.

8. Sector safety profiles

Safety means different things in an operating theatre, a power plant, and a classroom. We propose a common base for all sectors: the seven-layer stack, organisational management systems under ISO/IEC 42001, and system-level testing in accredited labs. On that base, each sector defines its own risk profile, tests, and thresholds, with regulators and industry leaders leading. National data protection authorities should co-lead wherever personal data drives decisions.

Table 5. Sector safety profiles

Sector	Why safety matters most here	Priority tests and controls	Who should lead the profile
Healthcare	Errors harm patients directly; data is highly sensitive; models trained elsewhere may miss local disease patterns	Clinical evaluation following the ITU, WHO, and WIPO Global Initiative on AI for Health; software quality under ISO 13485 and IEC 62304 where AI is a medical device; validation on local populations and in local languages; human sign-off on diagnosis; hallucination testing; privacy	Health ministries, medicines and device regulators, medical councils, data protection authorities, hospital groups
Industrial AI and energy	AI controls physical systems; failures can injure people or stop production	Safety-case testing on digital twins; fail-safe degradation rather than hard stops; OT cybersecurity; human override	Industry ministries, energy and mining regulators, national CERTs, manufacturers

Sector	Why safety matters most here	Priority tests and controls	Who should lead the profile
Financial services, including mobile money	Agents move money; fraud and bias scale fast; telecom operators run much of Africa's payments and micro-credit	KYA for every agent; fairness in credit, including thin-file scoring; fraud and prompt-injection red-teaming; pre-settlement controls; model risk management; stop-control drills	Central banks, jointly with communications regulators for mobile money; capital-markets and insurance regulators; data protection authorities; bankers' associations and mobile operators
Government and public services	Decisions affect rights; public trust is at stake	Impact assessment before use; explainability; appeal routes; procurement rules requiring evidence	Ministries of ICT and public service, schools of government, auditors general, data protection authorities
Education	Learners include minors; assessment fairness; skills for the future	Age-appropriate safeguards; fairness in grading; academic integrity; AI literacy in curricula	Ministries of education, examination councils, universities, data protection authorities
Transport and logistics	Agents route ambulances, fleets, and cross-border freight; errors cost lives and trade	Routing and dispatch testing under real conditions; fail-safe fallbacks; customs and border-clearance accuracy	Transport ministries and regulators, emergency services, revenue and customs authorities
Agriculture	Advice reaches smallholders who cannot verify it	Local-language accuracy, including in languages current benchmarks miss; testing on local crops and conditions	Agriculture ministries, research institutes, cooperatives
Telecoms and media	Scale of reach; deepfakes and misinformation	Content provenance; detection; abuse and hate-speech monitoring in African languages	Communications regulators (for Eswatini, ESCCOM), national CERTs, broadcasters

Industrial AI builds on UNIDO's work. UNIDO's Global Alliance on AI for Industry and Manufacturing (AIM Global) has set out Principles for the Responsible and Reliable Use of AI in Industry and Manufacturing, alongside regional hubs for industry and manufacturing in Europe, Africa, and Latin America (UNIDO AIM Global). The alliance now has more than 500 members in 70 countries and is moving from alliance-building to implementation (UNIDO, 2026). Its African chapter, the Alliance for Industry 4.0 and Smart Manufacturing in Africa (AISMA), brings this work to the continent (UNIDO, WSIS 2025). Our industrial safety profile should be grounded in these principles, and UNIDO, which has agreed to partner with ESWASA and AICE Africa on this work, is the natural co-lead for industrial test packs. Similarly, the healthcare profile should be grounded in the clinical evaluation framework of the ITU, WHO, and WIPO Global Initiative on AI for Health (ITU, WHO, and WIPO).

What African benchmarks can and cannot do yet. Masakhane's IrokoBench and AfroBench, along with the AfriMed-QA clinical benchmark, are strong, peer-reviewed foundations, but they measure capability,

not safety. AfriMed-QA is in English, and IrokoBench does not cover siSwati, Eswatini's national language. Safety-specific work is emerging: UbuntuGuard, published in 2026, is described by its authors as the first policy-based safety benchmark for African languages, built from adversarial queries written by 155 domain experts across 10 languages; AfriHate covers hate speech and abuse; MasakhaNER 2.0 supports detection and redaction of personal data; and INJONGO covers intent detection for conversational services such as banking. None of these yet tests multi-step agents, clinical tools in local languages, or adversarial tool use. Building those tests, starting with siSwati and other Southern African languages, is work the Eswatini test bed will take on with the research groups behind these datasets.

To make results comparable across sectors and borders, every profile should use a common scale of four operational risk tiers (minimal, limited, high, and critical). This African scale, proposed for development with ARSO and SADC regulators, determines how much control a permitted system needs; it maps to, but is not the same as, the EU AI Act's categories, and practices the law prohibits sit outside it altogether. Every profile should also answer the same six questions:

1. Which AI uses are high risk here?
2. What must be tested before deployment?
3. What must be monitored after?
4. Who may stop a system, and how?
5. What evidence will a regulator accept?
6. Who is liable, and how do people seek redress and appeal when a system causes harm despite being tested?

Two tests apply across every profile. Assurance must be proportionate, so that SMEs and services used in the informal economy, such as mobile-money agents and agricultural advice lines, can meet it at a cost they can bear. And testing must look for disparate impact on women, on people with low literacy or digital skills, and on speakers of languages that benchmarks miss, because they are often the users least able to detect or contest a failure.

9. Eswatini: the test bed the world needs

A small country with a strong national standards body and short lines between government, regulators, and industry is a practical place to demonstrate that AI safety can function as a national system. Eswatini is positioning itself as Southern Africa's reference point for trusted AI.

Where we are

The ESWASA and AICE Africa AI Safety and Standards Centre of Excellence was launched in Mbabane on 27 October 2026 (Capital FM; The Standard; APO via Africa.com). ESWASA provides institutional leadership, standards expertise, and national coordination. AICE Africa provides programme design, AI governance

expertise, platform development, and skills transfer. ESWASA staff and board have completed training in AI safety and standards. The lab is proposed and is seeking funding.

The Centre's work covers AI risk classification; ISO/IEC 42001 readiness; preparation of the proposed AI safety testing lab; SADC regulatory harmonisation; training AI ethics and safety officers, with certification to follow through an accredited scheme; and a citizen-protection layer. Its safety framework, announced at the launch, rests on five pillars, which map onto the layers of the stack in Section 2: risk classification, audit and assurance, safe operations, cybersecurity, and citizen protection.

Some of what the test lab will do

The proposed lab, working name Sentinel Digital, would be among the first in the region dedicated to evaluating AI systems using technical testing harnesses. It would operate independently of the Centre's advisory work and seek accreditation under ISO/IEC 17025. Its test reports would serve as evidence for certification decisions made by an accredited certification body and for the operational controls examined in ISO/IEC 42001 audits. The lab is designed to work in a closed loop with the deploying organisation, which runs monitoring and stop controls itself: test, monitor, observe, stop, and retest, with every result recorded on a tamper-evident evidence trail. An accredited test report covers the system configuration tested; keeping assurance current as models, data, tools, and permissions change is a separate, ongoing service (Section 12).

Table 6. Proposed functions of the Sentinel Digital lab

Lab function (working names)	What it does	Who it serves
Proving Ground	Sandboxed evaluation and red-teaming of a system before it goes live, producing a graded test report	AI vendors, CTOs, risk officers, procurement teams
Watchtower	Continuous monitoring for drift, changes in goals or access, and risks in open-weight models, published as an open reference specification and run inside the client's environment	CISOs, model risk teams, AI operations
Observatory	Downstream impact tracking, sector benchmarks, and an open-weights watch	Regulators, boards, buyers
Safeguard	Monitoring and graduated stop controls, published as an open reference specification and operated by the client inside its own perimeter, plus drills and incident replay	Risk officers and regulators of critical systems

For industry, the lab means a single place to obtain evidence that a system is safe enough to buy, deploy, or sell in African languages and conditions. **For regulators**, it means a technical arm that can verify claims rather than taking them on trust. **For Africa**, it means testing capacity that remains on the continent, builds local careers, and produces African benchmarks. ESWASA would own the national scheme and register, and operations would be handed over to Eswatini nationals over time. ESWASA is considering hosting the lab at a university once it is built.

Who holds the off switch? The lab never holds a live switch over a bank, hospital, or power grid. Monitoring and stop controls run within the client's perimeter, and the client retains physical and legal control of every stop decision, from throttling or suspending an agent to revoking its credentials. What leaves the client is limited to cryptographic audit proofs and metadata the client chooses to share. These proofs show that records have not been altered; they cannot prove that every event was recorded, so they are backed by independent configuration audits. A regulator can act in two ways: by requiring the client to stop a system under its existing statutory powers, or by triggering revocation of an agent's credentials at payment rails and gateways, which the client's own fail-safes then handle. Regulators should judge whether stop controls work, not require a particular product. Any direct external intervention would require an explicit legal mandate, fail-safe fallbacks, strong cybersecurity isolation, and clear liability rules, and this paper does not propose one.

Keeping advice, testing, and certification separate. AICE Africa provides programme design and platforms; the Centre of Excellence trains and supports readiness; the lab, under separate governance, tests; and an accredited certification body certifies. Accreditation standards (ISO/IEC 17021-1 and ISO/IEC 42006 for management-system certification, ISO/IEC 17025 for testing, and ISO/IEC 17065 for product certification) require each to remain impartial: no body certifies an organisation it has advised, and no lab tests a system it designed or configured. We will keep these functions under separate governance and decision-making, publish an impartiality policy, and never use the joint Centre name to market certification or suggest that readiness support leads to a certificate. The accreditation body will confirm in detail how these rules apply before accreditation is sought.

Before the lab opens. The lab targets initial operations in late 2027 and ISO/IEC 17025-accredited operations in 2028, but evidence is needed sooner. From early 2027 to mid-2027, we will run a pilot: cloud-hosted evaluations of two or three Eswatini enterprise or public-sector systems in isolated environments, using the Policy Studio prototype, Inspect, and AI Verify. Initial results will inform the SADC Convening in February 2027, and more comprehensive case studies will inform the Charter tabled at SHIELD 2027, ensuring regional commitments rest on real results.

Paying for it. The lab is designed to transition from donor and partner seed funding to self-sustaining revenue from fees for testing, subscriptions for continuous assurance, and training. Eswatini's domestic market alone cannot sustain an accredited AI lab, so its viability depends on serving clients across Southern Africa as a regional reference lab. That regional revenue is what would allow a host university to run the lab, and its operations to pass to Eswatini nationals, beyond the first grants.

The lab will be built on open standards where they exist, such as the UK AI Security Institute's Inspect framework (UK AISI Inspect), OpenTelemetry, and Singapore's AI Verify (AI Verify Foundation), and on African benchmarks such as IrokoBench, AfroBench, AfriMed-QA, and UbuntuGuard, respecting each dataset's licence, so that its evidence is portable and can be recognised by others. Evaluation compute will combine secure in-country capacity for confidential work with cloud capacity in African data centres, such as those in South Africa and Kenya, and the lab will seek access to the compute of the AI Hub for Sustainable Development for open benchmark work. Client models and data will be evaluated in isolated environments agreed with each client and used for no other purpose.

The Policy Studio: one test, many rulebooks

The Policy Studio maps laws and standards to technical tests and evidence. Each clause of an Act or standard becomes a control; when a control can be verified technically (robustness, security, prompt-injection resistance, statistical fairness), it is linked to an automated test that generates evidence. One technical control often supports several frameworks, so organisations can test once and reuse the evidence. However, many legal obligations cannot be reduced to automated tests. Lawful basis for processing, necessity and proportionality, adequacy of cross-border transfers, the competence of human oversight, and fundamental-rights impact assessments all require human legal review. The Policy Studio tracks these as documented workflows rather than pass-or-fail tests.

Its initial library covers Eswatini's data protection and cybercrime laws (both 2022); regional and continental instruments, namely the SADC Model Law on Data Protection, the AU Malabo Convention, and the AU Continental AI Strategy; data protection laws in key markets such as South Africa (Protection of Personal Information Act, POPIA), Kenya, and Nigeria (Nigeria Data Protection Act, NDPA); the EU AI Act and GDPR; the Council of Europe Framework Convention on Artificial Intelligence; the UNESCO Recommendation on the Ethics of AI; the NIST AI Risk Management Framework; ISO/IEC 42001, 23894, 42005, and 27090; the OWASP GenAI Top 10 and MITRE ATLAS; and national AI policies from countries such as Rwanda and Ghana. An Act simulator shows a regulator or company how far a system's technical controls and documentation cover another jurisdiction's requirements before it enters that market. It measures coverage and completeness; it does not predict how a regulator or court would exercise its discretion.

This is what opens markets, with one caveat. A fintech in Mbabane can use a single set of technical tests to build evidence for Eswatini, South Africa, Kenya, and the EU, rather than starting from scratch in each. That evidence speeds up regulatory submissions and buyer due diligence. It does not replace each jurisdiction's procedures: domestic filings and authorisations, or, for high-risk systems sold in the EU, the AI Act's conformity assessment, including notified-body review where required. For the EU, the Policy Studio produces a pre-conformity evidence package, not a CE mark.

The ambition is for Eswatini to anchor SADC and for sister centres to anchor other regions, sharing methods and recognising each other's results. We will build on and partner with existing initiatives: Kenya's membership in the International Network of AI Safety Institutes (European Commission), the African Hub on AI Safety, Peace and Security at the University of Cape Town (UCT), and the Masakhane research community.

10. Aligning with global standards, and trading AI across borders

Africa should neither copy other regions' rules wholesale nor create rules that no one else recognises. The goal is interoperability: African standards that reflect African priorities and map cleanly onto the frameworks used by our trading partners.

Table 7. Alignment with global and African frameworks

Framework	Status (November 2026)	How we align
AU Continental AI Strategy	Endorsed by the AU Executive Council in July 2024 (African Union); calls for technical standards to ensure the safety and security of AI systems	The Charter and the Centre give the strategy an implementing institution for safety, standards, and skills
ARSO and the ISO–ARSO Kigali Agreement	ARSO leads Africa's standards harmonisation; its 2025 Kigali Agreement with ISO strengthens Africa's role in international standards (ISO)	The route for a pan-African AI safety standard and for African input into ISO/IEC work
AU Malabo Convention and SADC Model Law on Data Protection	The continental and regional instruments for harmonising data protection and cybersecurity	Loaded into the Policy Studio as the regional baseline beneath national laws
EU AI Act	High-risk obligations apply from 2 December 2027 (Annex III) and 2 August 2028 (Annex I), as amended by Regulation (EU) 2026/1744 (EUR-Lex)	Policy Studio maps its articles to tests and documented workflows; African exporters get pre-conformity evidence that supports, but does not replace, EU conformity assessment
Council of Europe Framework Convention on AI (CETS No. 225)	Open for signature since September 2024, including to non-European states; binding only on states that ratify it (Council of Europe)	A comparative human-rights and rule-of-law reference for the Pan-African Charter
ISO/IEC 42001 and 42006	Published 2023 and 2025; adoption accelerating worldwide	The basis for certifying organisations' AI management systems
ISO/IEC 17025, 17065, and 17067	Established standards for testing labs and product certification; ISO/IEC 17067 revised in 2026	The basis for accrediting the lab's testing and designing system certification schemes
ISO/IEC AI testing and security standards	ISO/IEC TR 29119-11, 24029 series, TR 24027, TS 4213; ISO/IEC 27090 (2026)	Candidate methods for the lab's accredited scope and references for security testing
UNIDO AIM Global Principles	Principles for the responsible and reliable use of AI in industry and manufacturing (UNIDO AIM Global)	The starting point for the industrial AI safety profile and test packs
Linux Foundation open standards	Agentic AI Foundation (MCP, AGENTS.md, goose); A2A; OpenSSF Model Signing; Open Secure AI Alliance	The lab would test agents over these protocols, use model signing in bills of materials, and contribute African test packs back to open source
NIST AI RMF (United States)	Voluntary; widely used by US firms (NIST AI 100-1)	Mapped in the Policy Studio so US partners can read African results
UK and Singapore assurance tools	UK AISI Inspect; Singapore AI Verify	To be used in the lab so evidence is portable
UNESCO Recommendation on the Ethics of AI and Readiness Assessment Methodology	Adopted 2021 (UNESCO); readiness assessments under way in many African countries	Used as a national baseline for capacity building

Framework	Status (November 2026)	How we align
National laws (data protection, cybercrime, sector rules)	Vary by country	Each is loaded into the Policy Studio and tested alongside global frameworks

Africa as the test bed, not just the market

The world's standards bodies need places to test their standards under real conditions: lower-resource settings, many languages, mobile-first services, and agent-heavy finance. Africa offers all of these. When African labs test systems against both African and global standards, they generate evidence that improves the standards themselves. That makes Africa a co-author rather than a rule-taker.

Cross-border AI services

The AfCFTA Agreement has been signed by 54 of the 55 AU member states and ratified by 50 (tralac, 2026). AI services can cross its borders at the speed of software, but only if trust keeps pace. Three mechanisms make that possible:

1. **Accreditation and mutual recognition.** Once AI is added to the recognised scopes of SADCAS and its peers under AFRAC and Global Accreditation Cooperation Incorporated, test reports and certificates can be accepted in other member states, as ARSO already promotes for goods (ISO TBT). Each national regulator still decides what it relies on, and recognition by the EU or other partners would need formal agreements on top of this.
2. **Shared test packs.** A common core of security and structural tests can be run once and accepted everywhere, while language, clinical, and legal modules are run for each market, because a result in siSwati says little about performance in Amharic.
3. **Agent passports.** An agent registered and verified in one country can be checked anywhere it acts, through a common KYA schema.

Agent passports across 54 jurisdictions require careful design. Standards bodies such as ESWASA and its ARSO peers are well placed to govern the common schema and the list of trusted registries, while round-the-clock issuance and revocation of credentials are better handled by central banks, payment switches, and accredited trust-service providers, using open standards such as W3C Verifiable Credentials or verifiable Legal Entity Identifiers. Cross-border revocation, which spans chains of delegated agents, is still maturing (Section 12), so it should be piloted within SADC first. Liability must also be built in stages: from regulatory cooperation to reciprocal legal arrangements for enforcement and redress, and eventually to regional passporting, so that a deployer is answerable across participating markets without needing a legal representative in each.

With Europe, the United States, and other partners, the same mechanisms can open a two-way corridor: European and American AI services entering African markets with evidence African regulators accept, and African AI services entering theirs on the same terms.

11. Capacity development

Institutions are only as strong as the people in them. Africa needs thousands of people who can write, test, audit, and enforce AI safety, and it needs them before the rules bite, not after.

Table 8. Capacity development by audience

Audience	What they need	How we deliver it
Ministers, policy makers, and parliamentarians	Enough understanding to set strategy, fund institutions, and ask the right questions	Executive briefings; policy masterclasses; support for national AI strategies with a safety pillar
Regulators and standards bodies	Skills to classify risk, assess evidence, run sandboxes, and supervise AI	AI safety and standards training (already delivered to ESWASA staff and board); secondments to the lab; SADC peer exchange
Boards and executives	Oversight of AI risk and value	Enterprise AI workshops for CEOs, CTOs, and boards
AI safety officers and practitioners	Hands-on skills in testing, red-teaming, monitoring, and compliance	Training for AI Safety and Standards Officers, Ethics and Compliance Officers, and Monitoring and Technical Experts, grounded in ISO/IEC 17025 practice, published AI testing standards, and MITRE ATLAS, leading to accredited certification of people in time
Assessors and auditors	Accredited competence to certify and assess	Assessor pathway aligned to ISO/IEC 42001 and 42006, and to ISO/IEC 17025 for technical assessors
Students	Foundational AI literacy and a path into safety careers	A university programme giving students foundational AI knowledge; the lab as a training ground
Schools of government and public servants	Safe use of AI in public services	Safety programmes with schools of government
SMEs and local innovators	Safe-by-design skills and access to testing	Programmes through business associations, innovation hubs, and Business Eswatini
The public	Literacy to use AI well and spot harm	Public awareness campaigns and simple guidance

Programmes start from existing baselines rather than a blank page: UNESCO readiness assessments where countries have completed them, national AI strategies, and the experience of Africa's AI safety research hubs.

The model we favour is learning by doing. Over the past six years, AICE Africa has trained more than 2,500 technical professionals and more than 1,200 leaders through a Proof-of-Work approach grounded in real projects. The lab is designed to help local staff learn through hands-on evaluation work. Since clients entrust the lab with confidential models and data, trainees will work with open-weight models, synthetic systems, and anonymised cases. They will join live evaluations only with the client's consent. Seconded

officials will be subject to the same confidentiality requirements as staff and will not participate in evaluations of organisations supervised by their own regulator.

12. The future of AI safety: a decentralised, more capable optimiser

AI is moving from a few large models answering questions to vast numbers of agents acting on their own, built on open and closed models, running on clouds, phones, and edge devices. Safety will have to move with it: from checking a model once to governing a living network continuously. In this section, an 'optimiser' is any AI system that pursues goals; as such systems multiply and act for many different owners, optimisation itself becomes decentralised.

What changes as the optimiser decentralises

- **From models to networks.** Risk will come less from any single model and more from how many agents interact: agents negotiating with one another, buying from them, and delegating to them. Small drifts compound across a network.
- **From point-in-time to continuous.** An ISO/IEC 42001 certificate covers an organisation's ongoing governance processes, including change management. But a technical test of a specific system in January says little about that system once it has been updated, fine-tuned, and given new tools by June. An accredited test report covers what was tested; keeping assurance current requires change tracking, retesting when material changes occur, and periodic independent checks, offered as a continuing service alongside, not instead of, accredited testing and certification.
- **From inspection to identity.** When millions of agents act, inspecting each action is impossible. What scales is identity and authority: knowing which agent this is, who stands behind it, and what it may do. KYA becomes as basic to an AI economy as KYC is to banking.
- **From human-speed to machine-speed oversight.** Monitoring and stopping must happen in seconds. AI will increasingly be used to watch AI, with humans setting the rules and handling escalations. This creates its own risks: a monitored agent can embed instructions in its outputs to fool the monitor; monitors and agents built on the same model family can share blind spots; and floods of alerts can swamp human reviewers. Isolating a verifier model limits what it can do but not what it reads, so it can still be manipulated. Safe designs therefore validate and structure logs before a model sees them, use verifier models independent of the agents they watch, treat AI judgements as one signal among several, and keep deterministic, non-AI circuit breakers, such as hard rate and spending limits, authoritative for critical decisions.
- **From national to shared infrastructure.** No country can monitor a global agent network alone. Incident sharing, shared test packs, and mutual recognition become safety infrastructure in their own right.

Three horizons

Forecasts in AI date quickly, so we frame the future in terms of horizons rather than predictions. They describe what we expect the assurance agenda to require, not when specific capabilities will arrive.

Table 9. Three horizons for AI assurance

Horizon	What we expect to matter most	What assurance must be able to do
Now to 2028	Agents in payments, customer service, and public services; open-weight models fine-tuned locally; first AI rules taking effect	Test single systems before deployment; register agents; stop and revoke them; certify management systems; build evidence regulators accept
2028 to 2032	Networks of agents transacting with each other across borders; AI embedded in critical infrastructure	Assure systems continuously as they change; check agent identity across jurisdictions; share incidents and test results regionally
Beyond 2032	Systems that improve themselves or other AI; frontier capabilities governed by international agreement	Verify that safeguards survive modification; contribute African evaluators and evidence to international verification

The technical need

To meet this future, the safety stack must distinguish controls that can be deployed and audited today from capabilities that are still research problems.

Table 10. What can be deployed now, and what is still maturing

Deployable now	Still maturing
AI bills of materials and cryptographic model signing (for example OpenSSF Model Signing), tied to behavioural regression test reports from an accredited lab	Remote attestation that a modified open-weight model running elsewhere still has its safeguards (for example through trusted execution environments)
Agent credentials using W3C Verifiable Credentials or signed requests, with status lists and short-lived tokens, checked at gateways and payment rails within one ecosystem	Interoperable agent identity across many networks and jurisdictions, and revocation that spreads across registries and delegation chains at machine speed
Append-only, tamper-evident logs and transparency-log receipts shared across organisations, with raw data kept locally	Privacy-preserving audit trails that span many organisations end to end, and proof that logs are complete
Sandboxed evaluations of single systems against defined test packs and published test methods	Reliable evaluations of agents in long, multi-step tasks, detectors for subtle goal drift, and verifier models that resist manipulation

These are research problems, engineering problems, and business opportunities all at once.

The application layer

Most of the value, and most of the risk, will lie in the application layer: the agents and products that businesses and governments build on top of models. That is where African innovators can compete, because applications depend on local knowledge, languages, and relationships. It is also where safety must be built in through safe-by-design templates, sector-specific guardrails, and pre-launch testing.

The reconstruction of the web

We are living through a rebuild of the internet's foundations. Web 2 was built for humans clicking on pages; a growing share of traffic now comes from bots and agents. Web 3 technologies introduced ideas of verifiable identity, programmable money, and tamper-evident ledgers. The agentic web will combine both: agents with verifiable credentials, transacting with programmable payments, leaving auditable trails. To respect data protection law and local ownership of evidence, routine operation needs only cryptographic commitments, receipts, and credential-status proofs to cross organisational boundaries, while raw interaction data stays in sovereign local storage; when a dispute arises, authorised parties can be given controlled access to what they need.

For businesses, this means customer journeys, payments, marketing, and service will increasingly be carried out by and for agents. Firms that are not legible to agents, and cannot verify the agents that come to them, will lose business. For governments, it means public services, tax, and identity systems must be ready to deal with agents acting for citizens.

Why businesses and governments must use AI, not just regulate it

The biggest risk for Africa is not unsafe AI. It is being left out of AI. A government that does not use AI cannot regulate it well; a bank that does not deploy agents will lose customers to one that does. Safety is the way to go, not a reason to wait. The right posture is to adopt deliberately: start with well-chosen use cases, test and monitor them, and expand as evidence accumulates.

13. The opportunity stack for Africa

Safety creates value at every level, from national institutions to individual careers. We see six layers of opportunity, from the institutions that set the rules to the people who do the work.

Table 11. The opportunity stack

Opportunity layer	What it is	Who captures it
Institutions	National and regional safety centres, test labs, certification schemes	Standards bodies, regulators, universities
Infrastructure	Sovereign compute for evaluation, evidence systems, agent registries	Data-centre operators, telecoms, technology partners
Assurance services	Testing, red-teaming, monitoring, certification, incident response, insurance	Independent African assurance firms
Safety technology	Tools for KYA, drift detection, guardrails, African-language benchmarks	Local startups and research groups
Safe applications	Trusted AI products in health, finance, agriculture, education, and government	African innovators and enterprises
Skills and careers	AI safety officers, assessors, red-teamers, auditors	Graduates, professionals, public servants

For local innovators, the opening is clear. Global safety tools are rarely built for African languages, mobile money, informal economies, or low-bandwidth settings. Benchmarks for Swahili, isiZulu, siSwati, Yoruba, or Amharic; fraud tests for mobile-money agents; and safety evaluations for clinical tools on African populations are products the world needs and Africans are best placed to build. They can start from strong foundations, from Masakhane's IrokoBench, AfroBench, and the AfriMed-QA clinical benchmark to UbuntuGuard for culturally grounded safety, and extend them into agentic, sector-specific assurance.

For job creation, safety is labour-intensive in a good way. Red-teaming, evaluation, auditing, and monitoring require skilled people, and these jobs can be done here for both local and global clients.

For sovereignty, testing capacity on the continent means African institutions can assess, govern, and improve the systems they depend on, rather than accepting certificates issued elsewhere.

For investors and partners, safety infrastructure is a durable asset. Demand rises with every new AI deployment and every new regulation, as it has for cybersecurity.

14. Roadmap and calls to action

The path runs from a national centre in 2026 to a regional hub by 2029, with the SADC Convening and SHIELD 2027 as the two gates that carry the work beyond Eswatini.

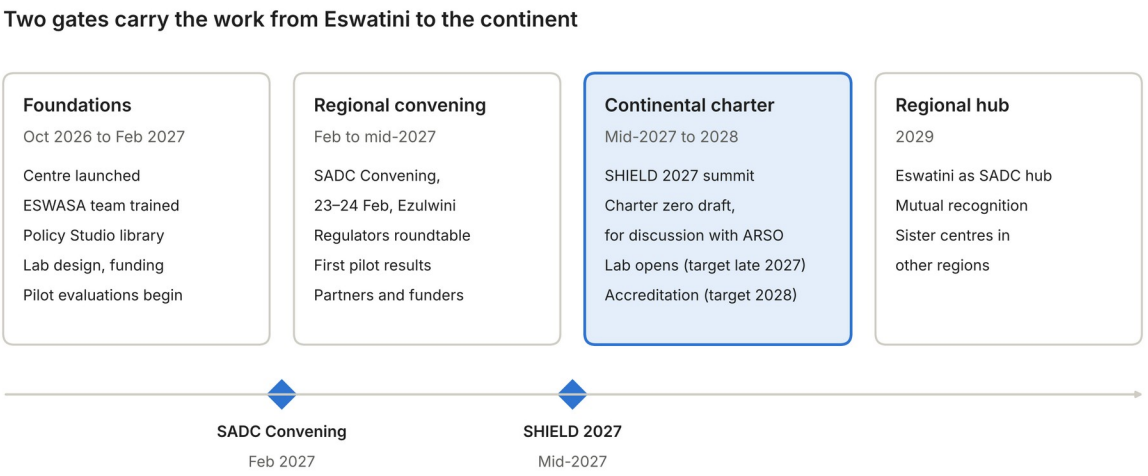


Figure 3. Programme roadmap. Two gates carry the work from Eswatini to the continent. Dates after February 2027 are targets.

The Centre of Excellence is in place, and the SADC Convening is scheduled. Everything after February 2027 is a target: the pilot evaluations run from early 2027 to mid-2027; the zero-draft Charter, developed for discussion with ARSO after the SADC Convening, will be tabled at SHIELD 2027; and the lab's opening in late 2027 and accreditation in 2028 depend on funding and on AI being added to recognised accreditation scopes. Accreditation of AI scopes is new ground: it needs validated test methods, proficiency testing or inter-laboratory comparisons, and trained technical assessors, and it depends on parallel progress in ISO/IEC JTC 1/SC 42, the joint ISO and IEC committee on AI, on AI test methods. Eswatini aims to be the pilot for SADCAS's first AI scope rather than a routine extension. The Charter phase is highlighted because it is where Eswatini's national work becomes a continental standard.

What we ask of each partner

Three asks matter most in 2027: fund the pilot evaluations and the testing lab; extend accreditation scopes to AI; and take part in the SADC Convening and SHIELD 2027. In full, we ask:

- **African governments:** put a safety pillar in every national AI strategy; plan AI safety capacity in every ministry and regulator; use AI in public services, with testing and monitoring from the start.
- **Regulators and standards bodies:** develop and adopt AI standards and sector safety profiles with industry; join the SADC Convening; work with ARSO towards mutual recognition of AI testing; consider statutory safe-harbour protection for payment rails, switches, and gateways that act in good faith on verified revocation signals from a recognised registry, with clear rules on who bears the loss from a mistaken revocation and how it is reversed.
- **Accreditation bodies (SADCAS, AFRAC, and peers):** extend recognised scopes to AI testing and AI management-system certification, and build a pool of AI technical assessors.
- **Enterprises:** appoint the three core AI safety roles (an AI Safety and Standards Officer, an AI Ethics and Compliance Officer, and an AI Monitoring and Technical Expert), alongside the technical leads

for agents, architecture, and red-teaming set out in Section 6; register every agent; test before you deploy and monitor after.

- **Model developers and platforms:** support independent testing in African languages and contexts; share safety evaluations; give deployers notice of model changes; offer audit access to accredited labs.
- **Global partners in the EU, the United States, and beyond:** recognise accredited African testing through formal agreements; co-develop standards and test packs; and make Africa a test bed for global standards.
- **Funders and development partners:** back the pilot evaluations, the testing lab, accreditation of AI scopes, the training programmes, and the SADC and SHIELD convenings.
- **The AfDB, UNDP, and the AI 10 Billion Initiative:** treat safety and assurance as part of AI readiness, so that the projects the initiative finances are tested, monitored, and trusted; connect the compute of the AI Hub for Sustainable Development to African evaluation work.
- **UNIDO and AIM Global:** co-develop the industrial AI safety profile and test packs on the AIM Global principles, through its Africa hub and the SADC work.
- **The Linux Foundation and the open-source community:** build KYA checks, revocation, and African test packs into open agent protocols and model-signing tools, so that safety travels with the software.
- **African innovators and universities:** build the safety tools, benchmarks, and services the continent and the world need, and train the people to run them.

Safety, done this way, does not hold back Africa's AI future. It is how that future becomes ours.

Acronyms

Acronym	Meaning
A2A	Agent2Agent protocol
AfCFTA	African Continental Free Trade Area
AfDB	African Development Bank
AFRAC	African Accreditation Cooperation
AICE Africa	AI Center of Excellence Africa
AIM Global	Global Alliance on AI for Industry and Manufacturing (UNIDO)
AISI	AI Security Institute (UK)
API	Application programming interface
ARSO	African Organisation for Standardisation
AU	African Union
CEO	Chief executive officer
CERT	Computer emergency response team
CISO	Chief information security officer
COO	Chief operating officer
CTO	Chief technology officer
DSIT	Department for Science, Innovation and Technology (UK)
ESCCOM	Eswatini Communications Commission
ESWASA	Eswatini Standards Authority
EU	European Union
GDPR	General Data Protection Regulation (EU)
Global ACI	Global Accreditation Cooperation Incorporated (formerly IAF and ILAC)
GPU	Graphics processing unit
IAF	International Accreditation Forum
ICT	Information and communication technology
ILAC	International Laboratory Accreditation Cooperation
IP	Intellectual property
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
IT	Information technology
ITU	International Telecommunication Union
JTC 1/SC 42	Joint Technical Committee 1, Subcommittee 42 of ISO and IEC (artificial intelligence)
KYA / KYC	Know Your Agent / Know Your Customer
LLM	Large language model
MCP	Model Context Protocol
MIRI	Machine Intelligence Research Institute
NDPA	Nigeria Data Protection Act
NIST AI RMF	US National Institute of Standards and Technology AI Risk Management Framework
OECD	Organisation for Economic Co-operation and Development
OpenSSF	Open Source Security Foundation

Acronym	Meaning
OT	Operational technology
OWASP	Open Worldwide Application Security Project
POPIA	Protection of Personal Information Act (South Africa)
REC	Regional economic community
SADC	Southern African Development Community
SADCAS	SADC Accreditation Service
SHIELD 2027	Africa AI Safety and Standards Summit, 2027
SME	Small and medium-sized enterprise
SOC	Security operations centre
UN	United Nations
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNIDO	United Nations Industrial Development Organization
W3C	World Wide Web Consortium
WHO	World Health Organization
WIPO	World Intellectual Property Organization

About the authors

AICE Africa (AI Center of Excellence Africa) is a pan-African company providing AI products and services, with a focus on AI sovereignty, safety, assurance, and ethics. It helps enterprises, governments, and institutions build, deploy, and operate AI, while strengthening their ability to control AI, assess its performance, and remain accountable for its use. The company operates in Kenya, South Africa, Nigeria, Tanzania, Eswatini, Mozambique, and Rwanda. It has trained more than 2,500 technical professionals and more than 1,200 leaders. Website: www.aiceafrica.com.

The Eswatini Standards Authority (ESWASA) is the national standards body of the Kingdom of Eswatini, responsible for standards development, conformity assessment, and quality infrastructure. Website: www.eswasa.co.sz.

Together they lead the AI Safety and Standards Centre of Excellence, launched in Mbabane in October 2026.

The Foreword is signed by John Kamara, Chief Executive Officer of AICE Africa (LinkedIn), and Ncamiso Mhlanga, Executive Director of ESWASA (LinkedIn).

Sources

Each inline citation matches one entry below. Standards named in the text are listed separately.

Official, intergovernmental, and legal documents

African Union (2024). Continental Artificial Intelligence Strategy. Cited as 'African Union'. https://au.int/en/node/44004

Council of Europe (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). Cited as 'Council of Europe'. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

European Commission (2024). First meeting of the International Network of AI Safety Institutes. Cited as 'European Commission'. https://digital-strategy.ec.europa.eu/en/news/first-meeting-international-network-ai-safety-institutes

European Commission (2026). Commission Implementing Decision (EU) 2026/2274 on the harmonised standards for high-risk AI systems. Official Journal, 14 October 2026. Cited as 'Implementing Decision (EU) 2026/2274'. https://eur-lex.europa.eu/eli/dec_impl/2026/2274/oj

European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Cited as 'EU AI Act' and 'Regulation (EU) 2024/1689'. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

European Union (2026). Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689 (Digital Omnibus on AI). Official Journal, 24 July 2026. Cited as 'Regulation (EU) 2026/1744' and 'EUR-Lex'. https://eur-lex.europa.eu/eli/reg/2026/1744/oj

International Organization for Standardization (2025). ISO and ARSO sign landmark Kigali Agreement. Cited as 'ISO'. https://www.iso.org/news/2025/10/iso-arso-kigali-agreement-trade

ISO WTO TBT Code register. ARSO: African Organisation for Standardisation, scope and mandate. Cited as 'ISO TBT'. https://tbtcode.iso.org/contents/json/194.html

ITU, WHO, and WIPO. Global Initiative on AI for Health. Cited as 'ITU, WHO, and WIPO'. https://www.itu.int/en/ITU-T/extcoop/ai4h/Pages/default.aspx

National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. doi:10.6028/NIST.AI.100-1. Cited as 'NIST AI 100-1'. https://doi.org/10.6028/NIST.AI.100-1

UK Department for Science, Innovation and Technology (2024). Assuring a Responsible Future for AI. Cited as 'UK DSIT, 2024'. https://www.gov.uk/government/publications/assuring-a-responsible-future-for-ai/assuring-a-responsible-future-for-ai

UK Department for Science, Innovation and Technology (2025). Trusted third-party AI assurance roadmap. Cited as 'UK DSIT, 2025'. https://www.gov.uk/government/publications/trusted-third-party-ai-assurance-roadmap/trusted-third-party-ai-assurance-roadmap

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. Cited as 'UNESCO'. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Accreditation

Global Accreditation Cooperation Incorporated (2026). Global Accreditation Cooperation Incorporated launch unifies international accreditation organisations. Cited as 'Global ACI, 2026'. https://iaf.nu/en/news/global-accreditation-cooperation-incorporated-launch-unifies-international-accreditation-organisations-and-strengthens-worldwide-trust/

International Accreditation Forum (2024). IAF General Assembly Resolution 2024-24: AB IAF MLA scope extensions. Cited as 'IAF, 2024'. https://iaf.nu/?p=5527

SADC (2015). SADC Accreditation Services (SADCAS) attains international recognition. Cited as 'SADC'. https://sadc.int/latest-news/sadc-accreditation-services-sadcas-attains-international-recognition

SADCAS (2019a). SADCAS achieves international recognition of its certification body accreditation scheme for management systems. Cited as 'SADCAS, 2019a'. https://www.sadcas.org/sites/default/files/2024-01/SADCAS_Achieves_International_Recognition_of_its_CBAS_-_MS.pdf

SADCAS (2019b). SADCAS maintains international recognition of its accreditation schemes. Cited as 'SADCAS, 2019b'. https://www.sadcas.org/sites/default/files/2024-01/SADCAS_Maintains_International_Recognition.pdf

Research and scientific reports

Abdullahi, T., Mgonzo, M., Oduwole, M., Okewunmi, P., Owodunni, A., Singh, R. and Eickhoff, C. (2026). UbuntuGuard: A Culturally-Grounded Policy Benchmark for Equitable AI Safety in African Languages. Findings of ACL 2026. Cited as 'UbuntuGuard'. https://aclanthology.org/2026.findings-acl.1663/

Adelani, D. I. et al. (2022). MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. EMNLP 2022. Cited as 'MasakhaNER 2.0'. https://aclanthology.org/2022.emnlp-main.298/

Adelani, D. I. et al. (2025). IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. NAACL 2025. Cited as 'IrokoBench'. https://arxiv.org/abs/2406.03368

Akerlof, G. A. (1970). 'The Market for "Lemons": Quality Uncertainty and the Market Mechanism'. *Quarterly Journal of Economics*, 84(3), 488–500. doi:10.2307/1879431. Cited as 'Akerlof, 1970'. https://doi.org/10.2307/1879431

Bengio, Y. et al. (2026). International AI Safety Report 2026. Cited by title. https://arxiv.org/abs/2602.21012

Muhammad, S. H. et al. (2025). AfriHate: A Multilingual Collection of Hate Speech and Abusive Language Datasets for African Languages. NAACL 2025. Cited as 'AfriHate'. https://aclanthology.org/2025.naacl-long.92/

Nimo, C., Olatunji, T. et al. (2025). AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset. ACL 2025. Cited as 'AfriMed-QA'. https://aclanthology.org/2025.acl-long.96

Ojo, J. et al. (2025). AfroBench: How Good are Large Language Models on African Languages? Findings of ACL 2025. Cited as 'AfroBench'. https://arxiv.org/html/2311.07978v5

Scher, A., Abecassis, D., Barnett, P. and Abeyta, B., MIRI Technical Governance Team (2025). An International Agreement to Prevent the Premature Creation of Artificial Superintelligence. Cited as 'MIRI'. https://techgov.intelligence.org/blog/new-report-an-international-agreement-to-prevent-the-premature-creation-of-artificial-superintelligence

Segun, S. T., Adams, R. et al. (2025). Toward an African Agenda for AI Safety. Cited by title. https://arxiv.org/pdf/2508.13179

Stanford Institute for Human-Centered AI (2026). AI Index Report 2026, Chapter 3: Responsible AI. Cited as 'Stanford AI Index 2026'. https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_3_responsible_ai.pdf

Yu, H. et al. (2025). INJONGO: A Multicultural Intent Detection and Slot-filling Dataset for 16 African Languages. ACL 2025. Cited as 'INJONGO'. https://aclanthology.org/2025.acl-long.464/

Yudkowsky, E. and Soares, N. (2025). *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. New York: Little, Brown and Company. Cited as 'Yudkowsky and Soares, 2025'.

Threat taxonomies, tools, and open specifications

AI Verify Foundation. AI Verify testing framework and toolkit. Cited as 'AI Verify Foundation'. https://aiverifyfoundation.sg/

MITRE Corporation. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. Cited as 'MITRE ATLAS'. https://atlas.mitre.org/

OpenSSF (2025). OpenSSF Model Signing (OMS) specification. Cited as 'OpenSSF, 2025'. https://github.com/ossf/model-signing-spec

OWASP Foundation. OWASP Top 10 for LLM and Generative AI Applications. Cited as 'OWASP GenAI'. https://genai.owasp.org/

UK AI Security Institute. Inspect. Cited as 'UK AISI Inspect'. https://inspect.aisi.org.uk/

Industry, legal, and market analysis

BSI (2026). BSI awards its 50th ISO/IEC 42001 certificate. Cited as 'BSI, 2026'. https://www.bsigroup.com/en-US/insights-and-media/media-center/press-releases/2026/november/bsi-awards-its-50th-iso-iec-42001-certificate/

CSO Online (2015). Worldwide cybersecurity market sizing and projections. Cited as 'CSO'. https://www.csoonline.com/article/2946017/worldwide-cybersecurity-market-sizing-and-projections.html

Cybersecurity Ventures (2025). Official 2026 Cybersecurity Market Report. Cited as 'Cybersecurity Ventures'. https://cybersecurityventures.com/official

ISS African Futures. Africa's current path. Institute for Security Studies. Cited as 'ISS African Futures'. https://futures.issafrica.org/thematic/01-africas-current-path/

MarketsandMarkets (2025). Cybersecurity Market Size, Share and Trends. Cited as 'MarketsandMarkets'. https://marketsandmarkets.com/Market-Reports/cyber-security-market-505.html

McKinsey & Company (2025). Leading, not lagging: Africa's gen AI opportunity. Cited as 'McKinsey'. https://www.mckinsey.com/ao/our-insights/leading-not-lagging-africas-gen-ai-opportunity

Microsoft Africa, reported by AllAfrica (2024). AI to add $1.2 trillion to Africa's economy by 2030. Africa's part of a $15.7 trillion global projection. Cited as 'Microsoft via AllAfrica'. https://allafrica.com/stories/202404110398.html

PYMNTS (2026a). 'The KYA Moment: Why Knowing Your Agent Is Becoming Table Stakes'. Cited as 'PYMNTS, 2026a'. https://www.pymnts.com/artificial-intelligence-2/2026/the-kya-moment-why-knowing-your-agent-is-becoming-table-stakes/

PYMNTS (2026b). 'Visa and Mastercard Team With Ant on Know Your Agent Framework'. Cited as 'PYMNTS, 2026b'. https://www.pymnts.com/cybersecurity/2026/visa-mastercard-team-with-ant-know-your-agent-framework/

SAP (2025). AI projected to add $1.5 trillion in Africa by 2030. Cited as 'SAP'. https://news.sap.com/africa/2025/06/ai-projected-to-add-15-trillion-in-africa-by-2030/

tralac (Trade Law Centre) (2026). Status of AfCFTA Ratification. Cited as 'tralac, 2026'. https://www.tralac.org/resources/infographic/13795

World Economic Forum (2025). How shared digital infrastructure can bridge the gap in Africa. Cited as 'World Economic Forum'. https://www.weforum.org/stories/2025/04/how-shared-digital-infrastructure-can-bridge-the-gap-in-africa/

Ecosystem initiatives

African Development Bank (2026). AfDB, UNDP and partners launch the AI 10 Billion Initiative during 2026 Nairobi AI Forum, via AllAfrica. Cited as 'AfDB via AllAfrica'. https://allafrica.com/stories/202602240717.html

ARC Advisory Group (2026). Open Secure AI Alliance finds a neutral home at the Linux Foundation. Cited as 'ARC Advisory Group'. https://www.arcweb.com/node/164037

InfoQ (2025). OpenAI and Anthropic donate AGENTS.md and Model Context Protocol to new Agentic AI Foundation. Cited as 'InfoQ'. https://www.infoq.com/news/2025/12/agentic-ai-foundation

Linux Foundation (2025). Linux Foundation announces the formation of the Agentic AI Foundation. Cited as 'Linux Foundation, 2025'. https://linuxfoundation.org/press/linux-foundation-announces-the-formation-of-the-agentic-ai-foundation

Linux Foundation (2026). Open Secure AI Alliance joins the Linux Foundation. Cited as 'Linux Foundation, 2026'. https://www.linuxfoundation.org/blog/open-secure-ai-alliance-joins-the-linux-foundation-to-build-a-shared-open-defense-stack-for-the-ai-era

TechAfrica News (2026). African Development Bank and UNDP advance $10 billion AI initiative to create 25 million jobs. Cited as 'TechAfrica News'. https://techafricanews.com/2026/10/16/african-development-bank-and-undp-advance-10-billion-ai-initiative-to-create-25-million-jobs/

UNDP (2026). Nairobi AI Forum 2026 drives AI adoption and impact. Cited as 'UNDP'. https://www.undp.org/node/569096

UNIDO. AIM Global: Alliance structure and principles. Cited as 'UNIDO AIM Global'. https://aim.unido.org/?p=88

UNIDO (2025). UNIDO contribution to WSIS 2025, AI for industry. Cited as 'UNIDO, WSIS 2025'. https://unctad.org/system/files/non-official-document/wsis2025_c19_unido_en.pdf

UNIDO (2026). AIM Global Action Forum 2026. Cited as 'UNIDO, 2026'. https://indico.un.org/event/1025691/

University of Cape Town (2025). UCT leads global conversation on AI safety. Cited as 'UCT'. https://www.news.uct.ac.za/article/-2025-10-02-uct-leads-global-conversation-on-ai-safety-peace-and-security

News and commentary

APO Group via Africa.com (2026). Eswatini Standards Authority and AICE Africa launch AI Safety and Standards Center of Excellence. Cited as 'APO via Africa.com'. https://africa.com/apo/eswatini-standards-authority-and-aice-africa-launch-ai-safety-and-standards-center-of-excellence

Becker, A. (2025). 'The Useful Idiots of AI Doomsaying'. *The Atlantic*. Syndicated copy. Cited as 'The Atlantic'. https://www.aol.com/news/useful-idiots-ai-doomsaying-113000871.html

Capital FM (2026). Eswatini, AICE Africa to establish AI safety centre to boost African standards capacity. Cited as 'Capital FM'. https://capitalfm.africa/eswatini-aice-africa-to-establish-ai-safety-centre-to-boost-african-standards-capacity/

Fortune (2026). Ex-Google engineer Nate Soares hasn't added to his 401(k) in a decade because he's an 'AI kills all humans guy'. 27 October 2026. Cited as 'Fortune'. https://fortune.com/2026/10/27/ai-humanity-nate-soares-401k-retirement/

Future of Life Institute (2025). Why Building Superintelligence Means Human Extinction (with Nate Soares). Cited as 'Future of Life Institute'. https://futureoflife.org/podcast/why-building-superintelligence-means-human-extinction-with-nate-soares/

Office of the Chair of the International AI Safety Report (2026). 2026 International AI Safety Report charts rapid changes and emerging risks, press release, 3 February 2026. Cited as 'International AI Safety Report 2026, Office of the Chair'. https://www.aap.com.au/aapreleases/cision20260203ae77099

The Standard (2026). Africa seeks control of AI safety as Eswatini launches standards centre. Cited as 'The Standard'. https://www.standardmedia.co.ke/sci-tech/article/2001561221/africa-seeks-control-of-ai-safety-as-eswatini-launches-standards-centre

The Tech (2026). Nate Soares makes the case against artificial superintelligence. Published April 2026, reporting a talk on 11 March 2026. Cited as 'The Tech'. https://thetech.com/2026/04/09/nate-soares-superintelligence

Standards named in the text

IEC 62304:2006+A1:2015. Medical device software – Software life cycle processes.

ISO 13485:2016. Medical devices – Quality management systems – Requirements for regulatory purposes.

ISO/IEC 17021-1:2015. Conformity assessment – Requirements for bodies providing audit and certification of management systems.

ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories.

ISO/IEC 17065:2012. Conformity assessment – Requirements for bodies certifying products, processes and services.

ISO/IEC 17067:2026. Conformity assessment – Fundamentals of and guidelines for conformity assessment schemes.

ISO/IEC 23894:2023. Artificial intelligence – Guidance on risk management.

ISO/IEC 24029 series (TR 24029-1:2021; 24029-2:2023). Artificial intelligence – Assessment of the robustness of neural networks.

ISO/IEC 27001:2022. Information security management systems – Requirements.

ISO/IEC 27090:2026. Cybersecurity – Artificial intelligence – Addressing security threats and compromises to AI systems.

ISO/IEC 42001:2023. Artificial intelligence – Management system.

ISO/IEC 42005:2025. Artificial intelligence – AI system impact assessment.

ISO/IEC 42006:2025. Artificial intelligence – Requirements for bodies providing audit and certification of AI management systems.

ISO/IEC TR 24027:2021. Artificial intelligence – Bias in AI systems and AI aided decision making.

ISO/IEC TR 29119-11:2020. Software testing – Guidelines on the testing of AI-based systems.

ISO/IEC TS 4213:2022. Artificial intelligence – Assessment of machine learning classification performance.