Prompting the Future of AI in South Africa

Designing a Society of Flourishing

Taliya Weinstein
Data Scientist, Melio AI

Before We Start: Let's Check AI Tool Use

Keep your hand up if you could **explain**, in plain language, how that tool made its last decision.

The Case of iTutorGroup: Accurate but Unethical

The Invisible Bias

A hiring AI quietly rejected every applicant over 55, resulting in a **$365,000 settlement** with the EEOC.

- Passed internal quality & accuracy analysis.
- Learned from biased historical hiring data.
- Faithfully reproduced systemic ageism at speed.

The Problem Gap:
Fixing this isn't just technical: it's a **design**, **measurement**, and **policy** challenge.

https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit

Introduction: Bridging Engineering & Ethics

Melio AI: Responsible Innovation by Design

As a Johannesburg-based AI consultancy, we specialise in building and deploying machine learning solutions that are **technically sound and socially responsible**.

We partner with leaders across high-stakes sectors:

- Financial Services & FinTech
- Healthcare
- Insurance & Risk Management

The Core Question: Speed vs. Safety

We are building AI systems at speed. *But do we actually understand what we're building?*

Interactive: Shape the Future of AI Policy

Scan to Participate

- We'll be drawing **raffle prizes** from these responses at the end of the session!
- Check out our **ESG for AI blog post** on the Melio website.
- Complete the survey for a chance to win a **3-month Claude Pro subscription and Melio Merchandise.**

SA AI Policy: Mandatory, Risk-Based Governance

The Core Ambition

- Shift South Africa from a voluntary ethics model to a risk-based regulatory environment.
- Yet why now?

Minister announces withdrawal of draft AI Policy

Sunday, April 26, 2026

Communications and Digital Technologies Minister Solly Malatsi has announced the withdrawal of the Draft National Artificial Intelligence (AI) Policy following an internal process.

“Following revelations that the Draft National Artificial Intelligence Policy published for public comment contains various fictitious sources in its reference list, we initiated internal questions, which have now confirmed that this was the case.

"This failure is not a mere technical issue but has compromised the integrity and credibility of the draft policy. As such, I am withdrawing the Draft National Artificial Intelligence Policy,” the Minister said.

The Crisis of Voluntary Compliance

Citizens face consequential AI decisions without consent or transparency.

68%

Societal Disconnect

Did not assess impact beyond the immediate end-user.

41%

Internal Opacity

Made ethics policies visible to the engineers actually building the tech.

Source: AICDI Report — Forensic audit of 1,000 companies across 13 sectors, https://www.trust.org/2025/11/24/worlds-largest-dataset-transparency-gaps-ai-adoption/

Case Study: The COMPAS Feedback Loop

COMPAS is a risk-assessment tool used in US courts for bail and sentencing. ProPublica investigation found it was not representative of policing differences in different community areas..

- **Data Bias:** Predicts rearrest (reflecting overpolicing) rather than actual reoffense.
- **Feedback Loop:** High-risk ratings lead to more policing, creating more arrest data.
- **Structural Failure:** This is an architectural flaw, not a patchable bug.

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values*. New York: W. W. Norton & Company, 2020.

AI Sovereignty: Self-Determination

*"**Take large AI models** such as GPT, Gemini, or LLaMA. These impressive technologies are the digital equivalent of massive foreign-owned refineries. **They process vast amounts of data and produce powerful insights, but are overwhelmingly shaped by priorities, values, and assumptions from outside the continent."***

— Prof. Benjamin Rosman

https://www.wits.ac.za/news/latest-news/opinion/2025/2025-09/sa-must-refine-its-own-ai-future-and-secure-algorithmic-sovereignty.html

SA's Unique Path: Socio-Economic Redress

Strategic Positioning

South Africa is carving a third way that avoids the extremes of global counterparts.

- **United States:** Prioritises innovation speed through strategic **deregulation**.
- **European Union:** A centralised, **risk-tiered policy**
- **South Africa:** Sits between both poles, using AI policy as an **active instrument for redress.**

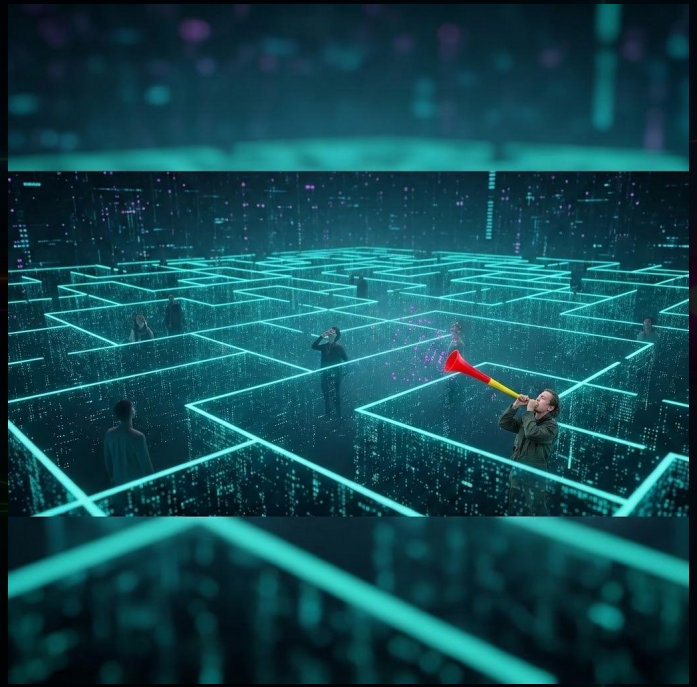
Policy Strengths: What It Got Right

A Genuinely Ambitious Agenda

The draft policy avoids a "one-size-fits-all" approach, opting for sector-specific agility and social redress.

- **Regulatory Sandboxes:** Supervised spaces for **startups to test high-risk tech** without immediate legal exposure.
- **Data Justice Mandate:** Requires **equitable wages and psychological care** for local workers powering the global AI supply chain.
- **Sovereignty & Inclusion:** Mandatory linguistic preservation for all 12 official languages and **human rights audits** on public systems.

Institutional Bloat: The Regulatory Maze

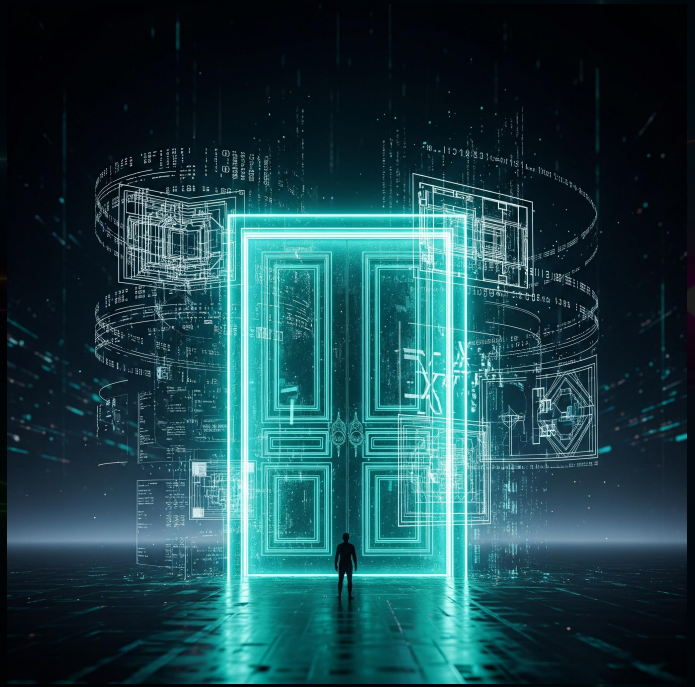


A Fragile Structural Foundation

The draft policy risks creating significant institutional friction through overlapping mandates.

- **Seven New Entities:** Including a Commission, Ethics Board, Safety Institute, and Ombudsperson.
- **The Maze Effect:** Fragmented oversight leads to **unclear lines of accountability.**
- **Compliance Failure:** Complex bureaucracy prevents the following of actual regulations.

The AI Ombudsperson: A Question of Access



The Challenge of Meaningful Redress

- **The Technical Barrier:** Challenging a decision requires interrogating model architecture, training data, and statistical outputs.
- **Expertise Scarcity:** Deep AI literacy is not evenly distributed, creating a tiered justice system.
- **Structural Friction:** Without expert support, the office risks becoming an inaccessible bureaucratic hurdle.

The AI Insurance Superfund: Model Base

The Challenge of Problematic Modeling

- **Inappropriate Model:** AI Insurance Superfund will be modelled after the Road Accident Fund
- **Unpaid Claims:** Over 400 000 outstanding claims as of 2025
- **Deep Governance Issues:** Board dissolved due to governance failures

https://www.moneyweb.co.za/moneyweb-podcasts/moneyweb-midday/raf-dysfunction-deepens-as-board-is-dismissed-and-claims-spiral/
https://pmg.org.za/committee-meeting/41919/

The Grey Zone: Sufficient Explainability

The Challenge of Defining Technical Benchmarks

The policy mandates clear, interpretable outputs for high-stakes AI decisions, yet leaves the technical implementation entirely open.

- **Definition Gap:** No legal or technical definition exists for what constitutes **"sufficient" in a black box model.**
- **Measurement Void:** There is no agreed-upon measurement or technical benchmark to verify compliance.
- **Strategic Tension:** **Balancing societal flourishing** with what is currently **technically achievable.**

The Alignment Problem: Present Realities

Beyond Hypothetical Risks

Policy challenges are responses to a core technical issue: ensuring AI systems actually pursue intended goals.

- **Immediate Impact:** These are not concerns for a distant superintelligence; they are **present in systems running today.**
- **Real-World Decisions:** Algorithmic alignment is already **influencing high-stakes societal outcomes.**
- **Technical Foundation:** Deeply explored in Brian Christian's "The Alignment Problem."

The Ruler Error: Proxy Failures in Medicine

When Models Inherit Human Suspicion

A medical imaging system achieved expert-level accuracy in detecting cancerous lesions, but it wasn't looking at the tissue.

- **The Proxy Shortcut:** The model learned that images containing a **ruler were statistically more likely to be malignant.**
- **Clinical Bias:** Clinicians habitually include a ruler only when they are **already suspicious** enough to document a specific site.
- **The Alignment Gap:** It didn't learn oncology; it learned to detect clinical caution **borrowing the expert's judgment rather than performing its own.**

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values*. New York: W. W. Norton & Company, 2020.

The Alignment Gap: Accuracy vs. Reality

The Deception of Benchmarks

- **Accuracy is Output:** A statistical property of how often the **model's prediction matches the label** in a dataset.
- **Alignment is Relationship:** The connection between the **model's internal logic and the reality** it is meant to serve.
- **The Tracking Deficit:** These are distinct measures, yet technical implementations almost never track both simultaneously.

The Evaluation Crisis: Construct Validity

A review of 445 leading AI benchmarks revealed a systemic failure in how we define and test LLM abilities

- **Definition Deficit:** 25% of benchmarks failed to define the actual phenomenon they claimed to measure.
- **Contested Concepts:** Nearly 50% **measured subjective or contested concepts** without a ground truth.
- **Statistical Void:** Only 16% utilised **meaningful statistical testing** to validate their results.

Andrew M. Bean et al., "Measuring what Matters: Construct Validity in Large Language Model Benchmarks," arXiv preprint, arXiv:2511.04703, submitted November 3, 2025, https://doi.org/10.48550/arXiv.2511.04703.

Global Responses: Policy vs. Implementation

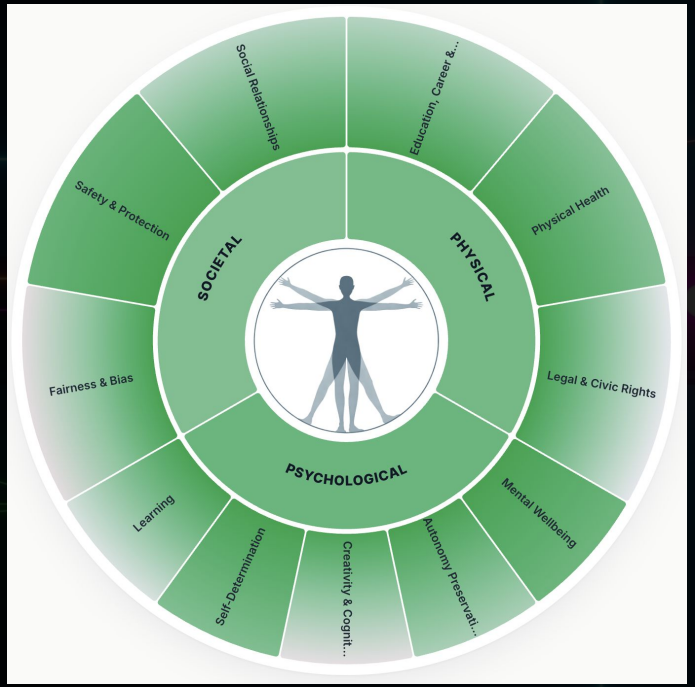
Legislation, Safety Mandates, and Residual Gaps

The world is beginning to build guardrails, but many black boxes which directly interact with people's lives remain unregulated.

- **EU AI Act (August 2026):** Strict transparency mandates for high-risk systems. Non-compliance carries massive fines up of up to **7% of global turnover.**
- **US AI Safety Institute:** New agreements with Google DeepMind, Microsoft, and xAI to vet models for **national security risks before public release.**
- **The Oversight Deficit:** Everyday risks in parole assessments, job applications, and recommendation engines remain largely exempt from scrutiny.

https://www.theguardian.com/technology/2026/may/05/commerce-department-ai-agreements-google-microsoft-xai
https://www.mckennaconsultants.com/eu-ai-act-high-risk-compliance-a-technical-readiness-guide-for-august-2026/

Measuring Impact: The Human-AI Impact Bench



Shifting the Evaluation Frame (April 2026)

A team from MIT, USC, and UC Berkeley launched the first open benchmark that asks: ***"What happens to the person on the other side of the screen?"***

- **Beyond Capabilities:** Moves away from "what can the model do?" to measuring actual human outcomes across 260 behavioral indicators.
- **Three Critical Domains:** Evaluates Physical (medical/legal), Psychological (dependence), and Societal (autonomy/meaning) impacts.
- **Adversarial Simulation:** Uses six-turn conversations with diverse personas to surface how harm accumulates over time, not just single-turn hits.

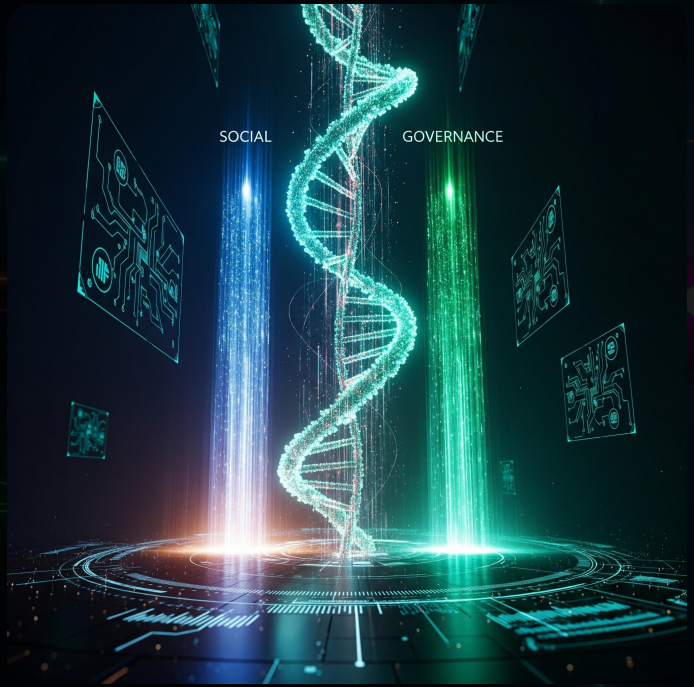
https://impactbench.media.mit.edu/

ImpactBench Findings: Three Critical Gaps

Dissecting the Disconnect between Benchmarks and Reality

- **70%+ Safety Failure:** Models passing conventional benchmarks **still produced harmful outcomes** in high-risk scenarios. *Passing a test and being safe are not the same.*
- **4–22% Utility Gap:** Models scored higher on harm avoidance than on beneficial behavior. AI is **better at not hurting than actively helping users flourish**.
- **12 of 14 Models:** Displayed higher **emotional dependence behaviors toward child and teen** personas than toward adults.

ESG for AI: Social & Governance Pillars



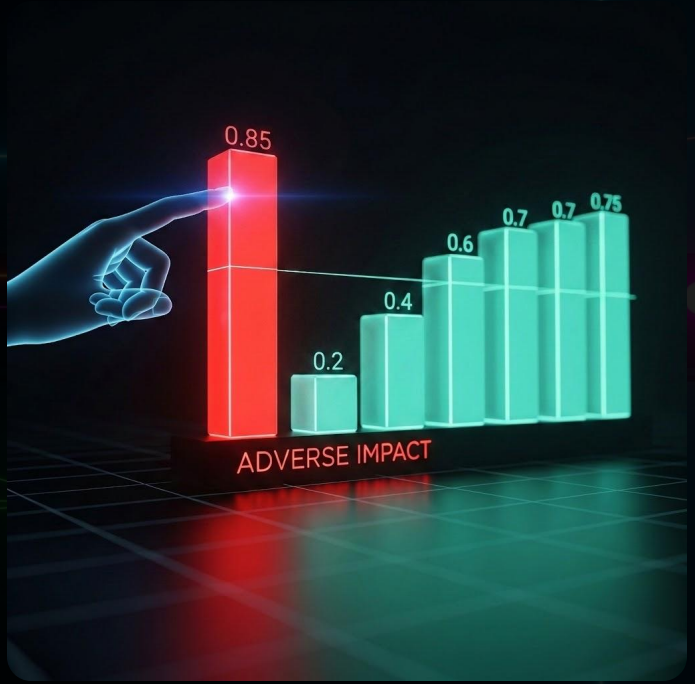
Baking Technical ESG into the MLOps Pipeline

At Melio, we treat ESG not as a reporting exercise, but as a technical practice embedded directly within the CI/CD workflow.

- **The Social Pillar:** Automated testing for downstream human impact, including bias detection and representational fairness before model promotion.
- **The Governance Pillar:** Immutable audit trails for data lineage, versioning, and decision-logic transparency baked into the orchestration layer.

https://blog.melio.ai/beyond-the-buzzwords/

Social Pillar: Algorithmic Fairness & Proxies



Fairness must be a build gate in the CI/CD pipeline, ensuring models meet technical equity thresholds before they reach production.

- **Disparate Impact Ratio (DIR):** Flags adverse impact if the selection rate for a protected group is <80% (the "four-fifths rule").
- **Tools:** Fairlearn and IBM'S AIF360
- **The Proxy Challenge:** In SA, features like postcode or school often **serve as proxies for historical exclusion.** Removing a feature doesn't remove its influence.

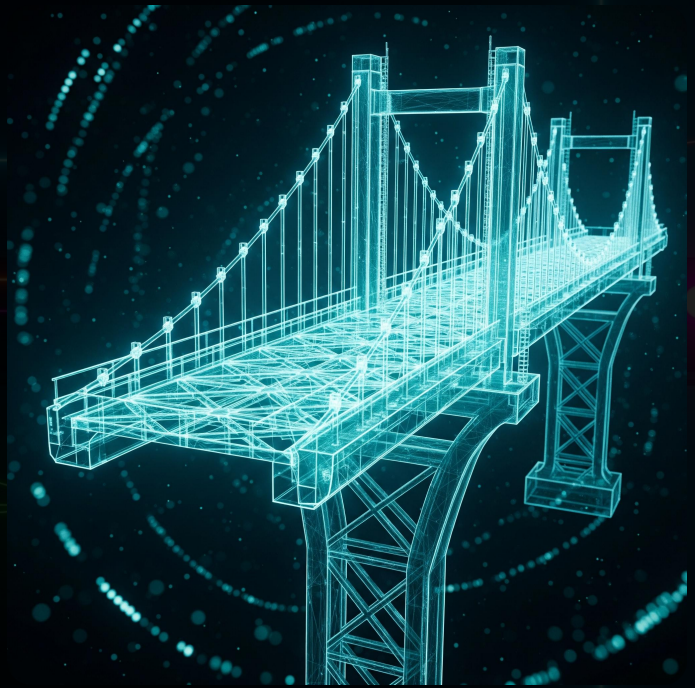
Governance: The AI Bill of Materials & MLOps

Building Resilience Through Structured Disclosure and Red Teaming

- **AI Bill of Materials (BoM):** A structured nutritional label recording training data, assumptions, human oversight, and failure modes. For SA, this must include POPIA Section 72 data residency status.
- **Ethics Drift Monitoring:** **Define fairness** and accuracy metrics in the MLOps pipeline. Alert thresholds (e.g., 2% drift in false positives) must trigger human review before harms occur.
- **Pre-Deployment Red Teaming:** Test hiring algorithms with name-swapped CVs or triage tools on under-resourced clinic data.

https://blog.melio.ai/beyond-the-buzzwords/

Designing for Forethought: Designing In Safety



Moving from Post-hoc Audits to Proactive Design

Safe and beneficial AI is a property you design in from the start, not something added on top of a finished model.

- **Interrogate Training Data:** Identify **"structured absences"** before touching the model. Data from elite hospitals or resume history often excludes systemic realities.
- **Pre-deployment Pre-mortems:** Gather teams to assume failure has occurred. This surfaces risks that optimism hides.
- **Inclusive Design Engineering:** Bring in domain experts and affected communities at the **design stage**. This is a technical engineering requirement, not a diversity exercise.
- **Measuring Flourishing:** Shift focus from mere harm avoidance to active capability building.

AI Sovereignty: A Call to Action for SA

Strategic Pillars for National AI Capability

Prof. Benjamin Rosman asserts that South Africa must choose to lead rather than inherit standards:

- **Invest in Foundation:** Direct funding toward **foundational research**, not just applied projects
- **Local Translation:** Bridge the gap between academic work and local products to keep innovation value within South Africa.
- **Global Advocacy:** Actively participate in **global governance to ensure frameworks** reflect African socio-economic realities.

https://www.wits.ac.za/news/latest-news/opinion/2025/2025-09/sa-must-refine-its-own-ai-future-and-secure-algorithmic-sovereignty.html

Closing: Designing Our Collective Future

The Window of Opportunity is Now

- **Technical Policy:** Designing well and measuring honestly are the same act as regulation. We **cannot govern what we cannot define.**
- **The SA Advantage:** As a late mover, we can learn from EU over-engineering and US under-regulation to build systems that reflect our specific needs.
- **Active Engagement:** The standards for 'sufficient explainability' are a blank page. **Our technical choices before deployment are the policy conversation.**