



Responsible AI by Design: Embedding Ethics Across the AI Research Lifecycle

ARISE KNOWLEDGE-SHARING WEBINAR SERIES #3

Event Report

Funder

The Andrew Carnegie Foundation

Principal Investigator

Prof. Nicholas Ozor, Executive Director
African Technology Policy Studies Network (ATPS)
P.O. Box 10081-00100, Nairobi, Kenya

Email: executivedirector@atpsnet.org | Telephone: +254 202714092

Website: <https://atpsnet.org/>

Date:

Wednesday, 23 September 2026

1. Executive Summary

The third webinar in the ARISE knowledge-sharing series, “Responsible AI by Design: Embedding Ethics Across the AI Research Lifecycle,” brought ARISE fellows and the wider AI African community together with two resource persons to work through what responsible AI requires in practice at each stage of a research project. The session preserved the concept note's three-part lifecycle structure — problem formulation and data collection, model development, and deployment.

The throughline across both presentations was consistent: technical soundness does not guarantee a responsible outcome. Bias, exclusion, and mistrust are introduced early, in how a problem is framed and whose voices inform the data, and compound through model development and deployment if left unaddressed. Both speakers grounded this argument in concrete, contrasting case studies and real-world incidents rather than abstract principles, in line with the practical, non-theoretical framing the session was designed around.

2. Event Details

Event Details	
Date	Wednesday, 23 September 2026
Duration	98 minutes
Format	Virtual webinar
Organizer	African Technology Policy Studies Network (ATPS)
Funder	The Andrew Carnegie Foundation
Project	ARISE — Advancing Responsible and Gender-Inclusive Artificial Intelligence Capacity in Africa's Higher Education Ecosystem
Principal Investigator	Prof. Nicholas Ozor, Executive Director, ATPS
Moderator	Azubuike N. Aniedu, ARISE Project Officer, ATPS
Resource Persons	Prof. Eric Tutu Tchao (KNUST / DIPPER Lab); Dr. Chukwuemeka Duru (Consumer Reports, USA)

Recording link: [View recording \(YouTube, 98 mins\)](#)

3. Attendance Note

Prof. Patrick Mensah (University of Energy and Natural Resources, Ghana) could not deliver the session on “Model deployment and accountability” as originally planned. Dr. Chukwuemeka Duru, Senior AI/Machine Learning Engineer at Consumer Reports, USA, stepped in and delivered a combined presentation covering both the model-development and deployment focal areas.

4. Welcome and Opening Remarks

Prof. Nicholas Ozor, Executive Director of ATPS and Principal Investigator of ARISE, opened the session by welcoming participants and reflecting on the pace of global debate around AI terminology and governance. He reaffirmed ARISE's mandate to build responsible AI capacity at the PhD and postdoctoral level across Africa, situating the project within the continent's broader need for doctoral-level AI/ML talent, and thanked the Andrew Carnegie Foundation for its continued support. He closed by encouraging the cohort to make full use of the webinar series and to apply themselves diligently in their research and fellowships.

5. Session Proceedings

5.1 Focal Area 1 — Laying the Ethical Groundwork: Problem Formulation and Data Collection

Speaker: Prof. Eric Tutu Tchao, Associate Professor and Head, Department of Computer Engineering, Kwame Nkrumah University of Science and Technology (KNUST); Scientific Director, DIPPER Lab.

Prof. Tchao's central argument was that responsible AI begins before writing a single line of code. Researchers, he observed, routinely fall into a “model-first trap”, that is, selecting an AI technique before the problem is properly understood, and this ordering, more than any downstream technical choice, is what produces biased, low-trust, low-adoption solutions. He offered a three-question problem-first framework that he urged every grantee to apply before starting a project:

- Whose problem is this? The affected community should help define the problem, not merely receive the solution.
- What does success look like? Success should be measured by real-world adoption and impact, not only by academic output such as publication or benchmark scores.
- Who bears the risk? — potential harms should be identified and mitigated in advance, including psychological or social risks to participants.

He illustrated the framework with two contrasting projects from his own research group:

	SIWA (Smart Indigenous Weather App)	AfroPalm
Problem origin	Defined with the affected community — rural farmers reliant on indigenous weather knowledge — through four years of embedded fieldwork by Dr. Enoch Bessa.	Defined independently by the researcher, based on a self-identified market problem (palm oil quality detection).
Data collection	Iterative and community-grounded; an early gender gap in the data — most participating farmers were men, though women were the primary end-users — was identified and corrected, raising model accuracy from roughly 72% in the first iteration.	Not community-grounded.

	SIWA (Smart Indigenous Weather App)	AfroPalm
Outcome	Over 500 active farmers using the tool, reaching a further 2000+ indirectly.	Published in a high-impact, well-cited journal, but never adopted by its intended users.

For the data collection stage specifically, Prof. Tchao proposed five checks he applies in his own work: whether consent was freely given; whether the data's origin is traceable; whether the sample genuinely reflects the population the system is meant to serve; whether communities are active participants in data production rather than passive subjects of observation; and whether the collection process builds local capacity rather than extracting value without benefit to the source community.

5.2 Focal Area 2 & 3 — Building Fair and Explainable Models, and Deploying with Accountability

Speaker: Dr. Chukwuemeka Duru, Senior Artificial Intelligence / Machine Learning Engineer, Consumer Reports, USA.

Where bias comes from

Dr. Duru framed AI bias as, in most cases, an unintentional consequence of training data rather than a deliberate design choice: models are a statistical reflection of the data they are trained on, and where a group is underrepresented in that data - as African languages, names, and locations are on much of the open internet - the model's performance for that group degrades accordingly. He illustrated the consequences with several documented real-world incidents:

Case	What happened	Consequence
Detroit Police facial recognition	A facial-recognition match error led to the wrongful detention of Robert Williams for roughly 30 hours.	Over US\$300,000 paid in settlement.
iTutor Group hiring software	The system systematically screened out qualified applicants aged 55 and above (ageism in an automated hiring pipeline).	Over US\$365,000 paid in settlement.
US commercial health-risk algorithm	Trained on historical healthcare expenditure, the algorithm under-predicted the healthcare needs of Black patients relative to white patients with comparable health status.	Correcting the bias raised the share of Black patients flagged for extra care from 17.7% to 46.5%.
US criminal recidivism-risk software	A long-used risk-assessment tool disproportionately rated Black defendants as high-risk for reoffending, reflecting historical socioeconomic patterns in the training data rather than individual risk.	Affected bail, sentencing, and release decisions before the tool was corrected.

He stressed that bias has to be checked at every stage of model development, not only at the point of data collection. During training, class imbalance across sub-groups causes the model's loss function to be dominated by the majority group, so minority-group performance degrades even as overall accuracy looks strong. At validation, he cautioned strongly against relying on a single blanket accuracy score, illustrating the point with a worked example: a Model A reporting the highest overall accuracy (91%) can still

perform far worse than a fairer Model C on a specific sub-group (62% for a rural population versus 84% under the fairer model) - a gap invisible unless performance is checked per group rather than in aggregate.

Deploying with accountability

On deployment, Dr. Duru recommended two concrete, low-cost mechanisms. First, structured feedback loops (as simple as a thumbs-up/thumbs-down signal) allow a team to detect data or user drift after launch; a rising share of negative feedback over time is an early warning sign that the model no longer matches how it is actually being used. Second, for African researchers who consume rather than train large pre-trained models (ChatGPT, Gemini, Claude, and similar), he recommended Retrieval-Augmented Generation (RAG) as a practical way to ground a model's responses in a curated, local knowledge base - for example, verified local health or agricultural information - constraining the model to answer only from that trusted source rather than its general training data, which reduces exposure to the model's underlying biases without requiring costly retraining.

On explainability, he noted an inherent trade-off: the more explainable a model is made, the less usable and the slower it typically becomes, so full explainability is rarely the right target for a production system. A practical reliability check he recommended is testing response consistency, that is, posing the same underlying question from different framings or perspectives and checking whether the model's answers remain coherent.

6. Moderated Discussion and Q&A

The extended Q&A session (25 minutes) drew active participation from ARISE fellows and external participants. Key exchanges are summarised below.

Topic	Raised by	Key response
Do I always need to train a new model?	Richard G. Wairagu, PhD Scholar	Both speakers cautioned against defaulting to heavier models. A simple model that meets the defined success criterion is sufficient; fine-tuning or knowledge distillation from an existing model is usually cheaper and more practical than training from scratch. Prof. Tchao also shared a self-evaluation framework for responsible AI development, available at facet.netlify.app .
How can AI reinforce racial inequality in the justice system?	Raised via chat	Dr. Duru cited the US recidivism-risk software case (above). Prof. Tchao added that Ghana is piloting AI-assisted legal research tools (Esquire, DIALegal) for lawyers and judges, and that human-in-the-loop review remains the essential safeguard against relying on biased or incorrect output in a live case.
How can the reliability of explainable-AI outputs be verified? Is there an accepted standard?	Dr. Ekle Francis Adoba, Postdoctoral Fellow, Obafemi Awolowo University	No single internationally accepted protocol exists. A practical check is to test whether a model's response stays consistent when the same question is posed in different framings or perspectives; this also surfaces the underlying explainability-versus-usability trade-off.

Topic	Raised by	Key response
Can a model still introduce gender bias if gender is not an explicit feature in the training data?	Mary Ofuru Kama, PhD Scholar, Ebonyi State University	Yes, bias can be encoded through variables that correlate with the missing attribute, even without it being present directly. This reinforces the case-study lesson from Focal Area 1, where excluding women from SIWA's early data collection skewed the model, even though gender was not the stated research variable.
What happens to data submitted to foreign-hosted models, and could it introduce bias through recursive self-improvement?	Dr. Barimwotubiri Ruyobeza, Postdoctoral Fellow, University of Cape Town	User data is typically stripped of personally identifying information before any reuse. Dr. Duru noted that African researchers largely consume rather than train these large models, which limits direct exposure to this risk but also limits control over how African data and languages are represented in the underlying model.
How should background noise or occlusion (e.g., in street-level image datasets) be handled?	Raised via chat	Dr. Duru pointed to established, lower-cost computer-vision techniques — OpenCV's image-processing suite, variational autoencoders, and diffusion-based reconstruction — as accessible tools for reconstructing or cleaning noisy image data without requiring expensive custom model training.

7. Key Takeaways for ARISE Grantees

- Start with the problem, not the model — involve the affected community in defining the problem and what success means before any technical work begins.
- Treat data collection as an ethical exercise: check consent, traceability, representativeness, community participation, and local capacity-building, not just data volume.
- Validate models using per-group metrics, not a single blanket accuracy score - aggregate performance can conceal poor outcomes for specific sub-groups.
- Bias mitigation is not a one-time step; it must be checked at data collection, during training, and again at validation.
- Where full retraining is impractical, fine-tuning, knowledge distillation, or RAG can ground a consumed model in local context at a fraction of the cost.
- Build feedback loops post-deployment (even simple thumbs-up/down signals) to catch model or user drift early.
- Treat explainability as a trade-off to be actively managed against usability, not a checkbox to be ticked once.

8. Closing Remarks and Vote of Thanks

Susan Mburu, ATPS Communications Officer, delivered the vote of thanks on behalf of the organizing team, thanking Prof. Ozor, the moderator, and the resource persons — including Prof. Patrick Mensah in absentia — as well as the Andrew Carnegie Foundation for its continued funding of the initiative. She reminded participants that this was the third edition of an ongoing webinar series and encouraged all attendees to complete the post-event feedback form circulated in the chat, to help shape future sessions.