GLOCAL EVALUATION WEEK 2026

Measuring What Matters

June 2, 2026 | 6:30 – 8:00 PM IST

Deepa Karthykeyan

Moderator
Co-Founder & Partner, Athena Infonomics

Dr. Francis Xavier Rathinam

Senior Director
Global MERL, Athena Infonomics

Aditi Namdeo

Independent Expert
AI for Social Good

Zeba Siddiqui

Senior Consultant
MERL, Athena Infonomics

AGENDA

90-Minute Webinar

01
Setting the Stage
Athena's AI vertical and the global context

02
Introduction to AI evaluation frameworks
Global AI governance and Athena’s framework

03
Technical Accuracy & Robustness
Pillar 1: Does the AI actually work where it counts?

04
Human-Context Validation
Pillar 2: Ethics, equity, and the people at the centre

05
Attributable Development Impact
Pillar 3: Who benefits, by how much, and at what cost?

06
Red Teaming
Pillar 4: Stress-testing for the real world

07
The AI Evaluation Ecosystem
Where Are We Now?

08
Synthesis & Takeaways
What you can use from today

09
Open Discussion
Questions from participants

What Prompted Athena's Framework

In conducting a landscape analysis of AI-enabled solutions in agriculture across L&MICs, Athena identified a critical gap: global AI governance frameworks exist but none are designed for development practitioners evaluating AI on the ground.

The Governance Gap

The Evidence Gap

The Practitioner Gap

CONTEXT

The Global AI Governance Landscape

Several landmark frameworks now govern the development and deployment of AI. They are consequential but were not designed for development practitioners working in resource-constrained contexts.

EU AI Act

2024

Risk-based regulation; compliance-heavy; oriented towards European markets and enterprise contexts

OECD AI Principles

2019/Updated

Five core principles: inclusive growth, human-centred values, transparency, robustness, accountability

UNESCO Ethics of AI

2021

First global standard; centres human rights, dignity, equity foundational to Athena's framework

China Generative AI Rules

2023

State-controlled content moderation; mandatory safety assessments before public deployment

US Executive Order on AI

2023

Safety, security and trustworthiness; mandates red-teaming for frontier AI models

India's AI Policy (NITI Aayog)

Evolving

AI for All emphasises inclusive growth, application to agriculture, health and education

The gap: these frameworks are complex, geography-specific, and do not translate easily for evaluators and programme managers working across L&MICs - which is where Athena's framework steps in.

INTRODUCTION

Athena's AI Evaluation Framework

A structured, four-pillar approach — simplified, replicable across contexts, and anchored in ethics and equity.

01

Technical Accuracy & Robustness

Does the model perform as intended in real-world, resource-constrained settings? Internal validity, causal accuracy, scalability.

02

Human-Context Validation

Are ethics and equity built into the design? End users as partners, not recipients. Reception, relevance, inclusivity.

03

Attributable Development Impact

Who benefits, by how much, and at what cost? RCTs, quasi-experimental methods, SDG-aligned metrics.

04

Red Teaming

Is the model stress-tested against real-world threats? Bias detection, data vulnerabilities, governance readiness.

PILLAR 01

Technical Accuracy & Robustness

What Do We Evaluate?

Prediction accuracy: How often does the model's output align with actual field outcomes yield, pest diagnosis, weather advisory?

Internal validity: Can we attribute outcomes to the AI model, accounting for confounding external factors?

Robustness: Does the model hold up under incomplete data, intermittent connectivity, or unusual conditions?

Scalability: Can the model be reliably replicated in different geographies and contexts with necessary customisation?

Causal accuracy: Are observed improvements truly driven by the AI, or would they have occurred anyway?

Why Internal Validity Cannot Be Assumed

- Data is heavily skewed toward large commercial farms. Smallholder farmer data although being the most critical context is systematically scarce.
- AI outcomes affect entire agricultural value chains. Errors in pest diagnosis or weather advisory cascade through livelihoods.
- An untested model cannot be scaled responsibly. Validation is the precondition for replication.
- In resource-constrained settings, model failure translates directly into crop loss and income shocks.

PILLAR 02

Human-Context Validation

> *"The pivotal idea is to advocate a stakeholder-driven approach. End users should have control over the development, design, and use of AI-enabled solutions to ensure responsible and sustainable outcomes."*

01

AI Lacks Contextual Understanding

However sophisticated the model, AI cannot capture emotional responses, cultural idioms, lived experience, or local agronomic knowledge without structured human input.

02

Reinforcing Patterns, Missing Innovation

AI excels at pattern replication. But locally adaptive innovation the kind that makes a solution stick in a specific community requires human values synthesised into design.

03

Increasing Uptake Through Insights

Human validation provides the critical appraisal that fine-tunes AI for granular, local realities. In labour-intensive smallholder agriculture, this is the mechanism of adoption.

04

The Continuous Learning Loop

Human context validation and AI work in tandem. A robust feedback loop compounds gains: each cycle produces a tool more aligned with community values and needs.

PILLAR 03

Attributable Development Impact

The central question: Who benefits from this AI intervention, by how much, and at what cost to whom ?

How AI Evaluation Differs from Traditional Impact Evaluation

Dimension	Traditional Evaluation	AI Evaluation
Nature of intervention	Static — defined at inception	Dynamic — evolves with data continuously
Primary focus	Attribution of cause and effect	Performance, adaptive behaviour, bias testing
Core methods	RCTs, DiD, Propensity Score Matching	A/B testing, algorithm audits, bias testing, red teaming
Design flexibility	Fixed once implemented	Continuous, real-time, and iterative
Complexity dimension	Human behaviour and context	Human–technological interface + ethical overlay

PILLAR 04

Red Teaming

Red teaming is a critical component for testing powerful AI systems - a systematic process to proactively counter vulnerabilities by emulating the techniques that could be used by bad actors."

What It Means for Development AI

AI Safety

Identifies harmful or unintended outputs before they reach vulnerable users including data manipulation, model hallucinations, and failure cascades under field conditions.

Bias Reduction

Stress-tests for differential treatment across gender, geography, age, and income the most common failure modes in agricultural AI deployed in L&MICs.

Reliability Under Adversity

Verifies model accuracy with unusual weather data, ambiguous satellite imagery, incomplete records conditions that are routine, not exceptional, in the field.

Transparency & Accountability

Lack of trust is the primary barrier to AI uptake. Red teaming generates the audit trail that enables regulators, funders, and communities to trust the system.

Responsible AI by Design

Catching biased outputs early before deployment is both ethically imperative and economically rational. Retrofitting fairness is far more costly than building it in.

The AI Evaluation Ecosystem — Where Are We Now?

How AI is Currently Evaluated

Evaluation happens primarily at development stage accuracy benchmarks and technical audits dominate

Most tools are assessed in controlled environments, not the field conditions they are actually deployed in

Post-deployment evaluation is rare and rarely systematic

Key Gaps

Evaluation is rarely participatory — communities and frontline practitioners are absent from the process

Governance frameworks are fragmented and rarely sector-specific, making cross-context comparison difficult

Equity dimensions — gender, digital access, language — are treated as add-ons, not design requirements

What's Missing

No consistent framework spans the full lifecycle from development to deployment to impact

Social, cultural and contextual factors are systematically excluded from evaluation design

Shared accountability between developers, funders and communities does not yet exist

Building Evaluation Models That Work for People

01

Evaluation frameworks built with communities not for them. End users and frontline practitioners shape what gets measured, how it is measured, and what counts as success.

02

Moving from one-time assessments to embedded evaluation loops across the full AI lifecycle development, piloting, deployment, and scale. Evaluation must keep pace with the technology.

03

Shared evaluation models across developers, implementers, funders, and practitioners with Athena's four-pillar framework as a common language for responsible AI across contexts.

Global Evaluation Initiative

What You Can Apply from Today

Global Evaluation Initiative

Open Discussion

Glocal Evaluation Week 2026

www.glocalevalweek.org

Global Evaluation Initiative

Thank you!

Glocal Evaluation Week 2026

www.glocalevalweek.org