WITNESS
SEE IT FILM IT CHANGE IT

# Response to NIST's AI Standards "Zero Drafts" Pilot Project

SUBMITTED ON: 30 MAY, 2025
BY WITNESS

This submission aims to provide input to NIST's AI Standards "Zero Drafts" Pilot Project, in light of the current call for responses. The present document addresses the *"Technical Measures for Reducing Risks Posed by Synthetic Content"* category and its subsection 'b', namely "Methods and Metrics for Evaluating and Reporting the Effectiveness of Such Measures."

***

WITNESS appreciates NIST's ongoing efforts to accelerate the development of trustworthy AI standards. In response to the "Zero Drafts" call for input, we strongly advocate for the integration of sociotechnical evaluation frameworks—such as WITNESS' Truly Innovative and Effective AI Detection (TRIED) Benchmark—into the development of standards for AI-generated or AI-manipulated content detection. These frameworks are essential to accurately assess and reduce the real-world risks posed by synthetic content.

Current detection metrics often over-index on technical performance, overlooking critical issues of usability, accessibility, relevancy, and capacity for innovation in addressing the challenges in deployment. AI detection tools must be evaluated in the context of diverse global realities, particularly for those directly engaging with deceptive synthetic media—such as journalists, fact-checkers, human rights defenders, civil society actors, and marginalized communities. Standards must be shaped not only by technical developers but also by those operating in high-risk information ecosystems to account for challenges limiting the usability of detection tools stemming from real-world contexts.

The *TRIED Benchmark* was developed through a multi-stakeholder process informed by over two years of practical implementation via the Deepfakes Rapid Response Force (DRRF), groundbreaking initiative connecting frontline information actors with leading media forensics and deepfakes experts to deliver timely evidence-based analysis of suspected deceptive AI content. It reflects lived experience and global input from partners in India, Sudan, Mexico, Georgia, Ghana, and beyond. The Benchmark offers a robust, field-tested framework for evaluating AI detection tools beyond conventional accuracy metrics.

## 1. Key Principles for Reducing Risks Associated with AI Detection Tools

Technical performance alone is an insufficient measure for AI detection effectiveness. While AI detection tools have the potential to provide crucial real-time support in high-stakes situations, in practice, they often fall short due to factors such as gaps in training data or technical constraints from compressed or low-quality media. Real-world deployment demands a sociotechnical perspective, which involves an innovative and inclusive analysis of how tools operate across

varied social, cultural, linguistic, and political contexts. *TRIED Benchmark* expands conventional approach to evaluation beyond algorithmic accuracy to include usability, relevance, and transparency.

The sociotechnical approach reflects emerging global policy norms. The EU AI Act, the National Institute of Standards and Technology (NIST), and the Organisation for Economic Co-operation and Development (OECD) have all emphasized the importance of trustworthy, human-centric AI, with transparency, robustness, and fairness as core principles. *TRIED Benchmark* aligns with these frameworks by offering actionable measures to implement these values in the context of AI detection. Through proposed mechanisms, the framework bridges the gap between ethical AI commitments and their real-world application, supporting the development and deployment of responsible and innovative AI detection solutions.

Based on WITNESS' global consultations and the DRRF's work—the TRIED Benchmark proposes six key pillars to guide evaluation of AI detection tools:

1. **Performance in Real-World Conditions:** Detection tools must be tested on media typical of real-world environments: low-resolution, compressed, multilingual, dynamically edited, and noisy content found on messaging platforms and social media. Evaluation metrics should account for resilience to compression artifacts, degradation, and contextual variation.
2. **Transparency and Explainability:** Detection outputs must be interpretable by non-technical users. Tools should clearly communicate their intended purpose, capabilities, and limitations. Evaluation metrics should assess how well tools support public trust, responsible AI literacy, and informed decision-making by journalists and fact-checkers.
3. **Targeted Accessibility and Usability:** Evaluation must include whether tools are accessible to their target users—particularly in low-connectivity or resource-limited settings. Usability should be measured in terms of interface design, language support, affordability, and adaptability across skill levels.
4. **Fairness and Representation:** Metrics must evaluate dataset and input diversity across demographic groups, languages, geographies, and media types. Fairness in training data is foundational, as it directly impacts detection accuracy and equity. Tools must be tested across varied inputs that reflect the lived realities and expertise of global users to avoid disproportionate failure rates among marginalized populations.
5. **Durability and Resilience:** Standards should require regular evaluation and updates to reflect new generative techniques and adversarial tactics. These evaluations should include input from external teams and diverse stakeholders. Durability metrics should evaluate tool maintenance practices, update frequency, retirement processes, and resilience over time.
6. **Integration into Broader Verification Ecosystems:** Detection tools must be assessed as part of broader verification workflows. AI tools alone do not establish authenticity; rather, they contribute evidence that must be contextualized through open-source

investigation, human oversight, and metadata analysis. Evaluation should measure how effectively tools integrate into these multi-layered verification processes.

**2. Policy and Standards Recommendations**

We urge NIST and relevant standards bodies to integrate these principles into standards development and advocacy to ensure that synthetic content risk mitigation tools are resilient and globally relevant:

- **Adopt evaluation frameworks like the TRIED Benchmark** that embed sociotechnical considerations into formal standards for AI detection tools.
- **Establish minimum effectiveness requirements** that reflect real-world challenges such as poor media quality, linguistic diversity, and usability barriers.
- **Mandate transparency metrics** for explainability, including clear disclosure of intended use and tool limitations.
- **Require fairness audits** and representative testing datasets as a condition for standards compliance.
- **Support global, multi-stakeholder collaboration** in developing, testing and validating AI detection tools and benchmarks.

***

**About WITNESS**

WITNESS is a global human rights organization that empowers people to use video and emerging technology to defend and protect human rights. Working across five regions—Asia and the Pacific, Latin America and the Caribbean, the Middle East and North Africa, Sub-Saharan Africa, and the United States—we collaborate with those most excluded or at risk, identifying gaps, designing solutions, and co-developing strategies to hold the powerful accountable and drive lasting change. We respond to critical situations by equipping affected communities with essential skills in audiovisual AI and video production, safe and ethical filming techniques, and advocacy strategies.

This submission was prepared by the Technology Threats and Opportunities (TTO) program, which scales our global community work at a systems level—sharing insights across regions, collaborating with diverse stakeholders with both lived experience and professional expertise, connecting communities facing similar challenges, and advocating for grassroots perspectives in technology and policy spaces. The program proactively engages with emerging technologies that shape trust in audiovisual content, ensuring they are developed and deployed in ways that protect, rather than undermine, human rights.