WITNESS

SEE IT FILM IT CHANGE IT

Overview of TRIED: WITNESS' Truly Innovative and Effective AI Detection Benchmark

INTRODUCTION

The proliferation of generative AI and the increasing ability to create deceptive synthetic media escalates the threat to public trust and information credibility, particularly in resource-constrained regions of the Global Majority. As information circulates on social media and communication platforms at an unprecedented speed and scale, existing content moderation cannot effectively address the challenge of deceptive synthetic media. In response, information actors increasingly turn to detection tools. Yet, these tools frequently fail to deliver reliable results in the global high-stakes, real-world environments where they are most needed.

To address the gap between the technical capabilities and practical applications of detection tools, WITNESS[1] introduces the ***Truly Innovative and Effective AI Detection (TRIED) Benchmark***[2]—a framework that provides a structured approach for evaluating the effectiveness of AI detection tools through innovative technical and sociotechnical lenses. Grounded in real-world cases of deceptive AI and informed by global consultations, *TRIED Benchmark* offers actionable guidance for developers, policy actors, and standards bodies to design and assess accountable, transparent, and user-centered detection solutions.

This brief provides an overview of ***TRIED Benchmark***, sharing its key findings and recommendations. It brings together real-world examples of deceptive AI to illustrate the different dimensions of effectiveness discussed throughout the report, emphasizing the need for a holistic assessment of AI detection tools accounting for their real-world functionality and prioritizing adaptability, transparency, accessibility, contextual relevance, and fairness. By embedding these considerations into the evaluation of AI detection technologies, *TRIED Benchmark* helps align the development of detection tools with broader policy goals that effectively tackle synthetic media in diverse, real-world environments and drive responsible, human-centric AI innovation[3].

Read the full version of our report, TRIED: Truly Innovative and Effective AI Detection Benchmark:

(Click or scan the QR code)

Read more on WITNESS' AI work:

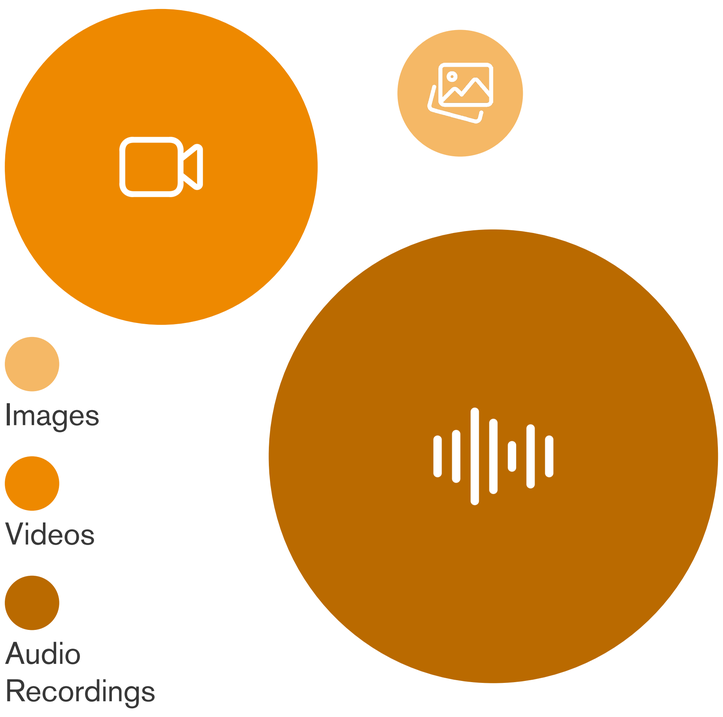
(Click or scan the QR code)

EVIDENCE FROM THE FRONTLINES: WHY DETECTION MATTERS

Since 2018, WITNESS has been leading the 'Prepare, Don't Panic' initiative on synthetic media, deepfakes, and multimodal generative AI[4]. In response to the lack of access to reliable and transparent detection tools[5] identified through our global work[6], in March 2023, WITNESS launched the Deepfakes Rapid Response Force (DRRF)[7], a pioneering initiative that connects frontline actors with media forensic and deepfakes experts to deliver evidence-based, timely analysis of suspected AI content. Over the past two years, the DRRF has supported information actors across India, Mexico, Ghana, Sudan, Ukraine, Venezuela, and Georgia, spanning audio, video, and images in both electoral and conflict contexts.

30% of analysed cases were received from partners in Africa, 28% from Asia, 28% from Europe and 14% from Latin America.

This global real-world engagement has revealed the dual potential and limitations of detection technologies[8]. DRRF interventions have helped defuse crises, build media literacy, and strengthen public awareness of synthetic media threats. However, the cases we have analyzed also expose persistent barriers to effective detection: underperformance with content involving high compression and low resolution, challenges with non-dominant languages and global public figures, a lack of explainability, and widespread difficulties in interpreting detection outputs due to limited AI media literacy. Compounding these technical limitations is a rising trend of false claims that real media is AI-generated[9], which further undermines public trust.



Out of all cases analysed, 64% were audio recordings, 31% were videos and 5% were images.

The inadequacy of many AI detection tools in practical use not only hampers their effectiveness but also heightens human rights risks. Infrequent tool updates, lack of transparency, and exclusionary design practices restrict their accessibility and reliability, especially for marginalized communities. Without equitable, context-aware, and resilient detection systems, synthetic media will continue to erode information credibility, amplify the human rights risks posed by deceptive AI[10], and deepen global disparities.

AN INNOVATIVE AND EFFECTIVE SOLUTION: THE TRIED BENCHMARK

Technical performance alone is an insufficient measure for AI detection effectiveness. 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.

This approach reflects emerging global policy norms. The EU AI Act[11], the National Institute of Standards and Technology (NIST)[12], and the Organisation for Economic Co-operation and Development (OECD)[13] 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 technologies.

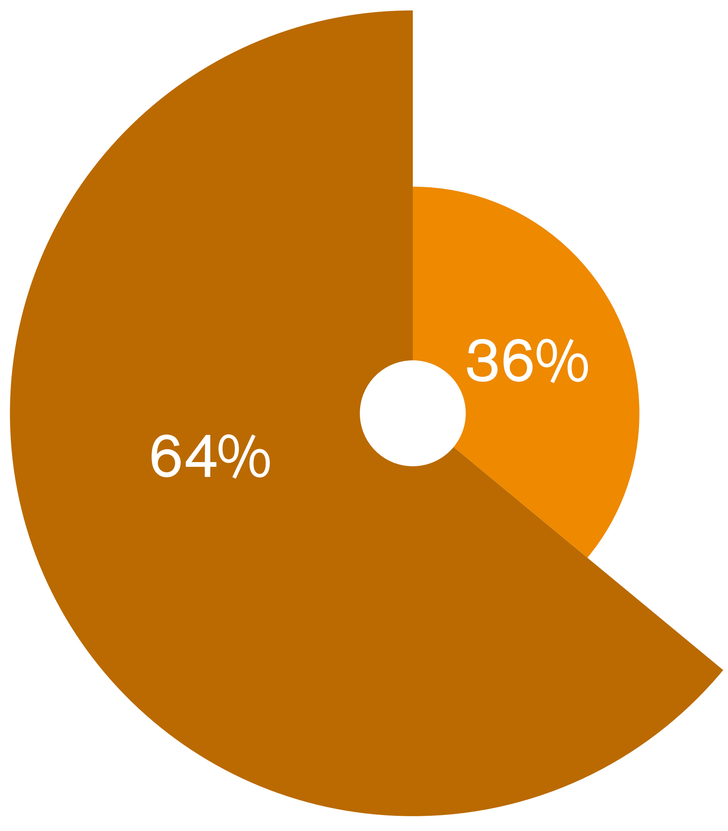
Additionally, the UN Guiding Principles on Business and Human Rights[14] and the revised OECD Guidelines for Multinational Enterprises on Responsible Business Conduct[15] provide a strong basis for embedding the *TRIED Benchmark* into due diligence processes and facilitate responsible AI detection through the Business and Human Rights framework. The benchmark helps assess adverse impacts on affected users, foster inclusive evaluation metrics, and guide mitigation practices grounded in meaningful collaboration between diverse stakeholders. We urge both States and industry actors to prioritize a multistakeholder approach in the development of technologies and policies that encourage global collaborations with local enterprises, and foster accessible and community-led technical solutions that are aligned with real-world challenges threatening information credibility.

REDEFINING INNOVATIVE EFFECTIVENESS OF AI DETECTION TOOLS IN PRACTICE

The **TRIED Benchmark** outlines six key elements for building and evaluating truly effective and innovative AI detection tools. These interconnected considerations stem from extensive real-world use cases gathered through the work of the Deepfakes Rapid Response Force, community feedback, and experts' consultations led by WITNESS, and reflect the practical needs of fact-checkers, journalists, and human rights defenders.

1. **Performance in Real-World Conditions:** Detection tools must be designed to handle the complexity and variety of real-world audiovisual content. For example, when analyzing a leaked conversation between key Nigerian public figures[16], the forensic experts couldn't reach a conclusive decision using AI detection tools because of low quality, high compression, and background music. While many tools perform well on high-resolution, near-original files, this does not reflect the reality of most media encountered by information actors. Content often analyzed by fact-checkers, journalists and human rights actors is diverse, dynamic, noisy, and compressed social media and messaging apps. It is key that common features and trends (such as platforms' compression or audio background noise) should not affect the tool's ability to provide reliable results.

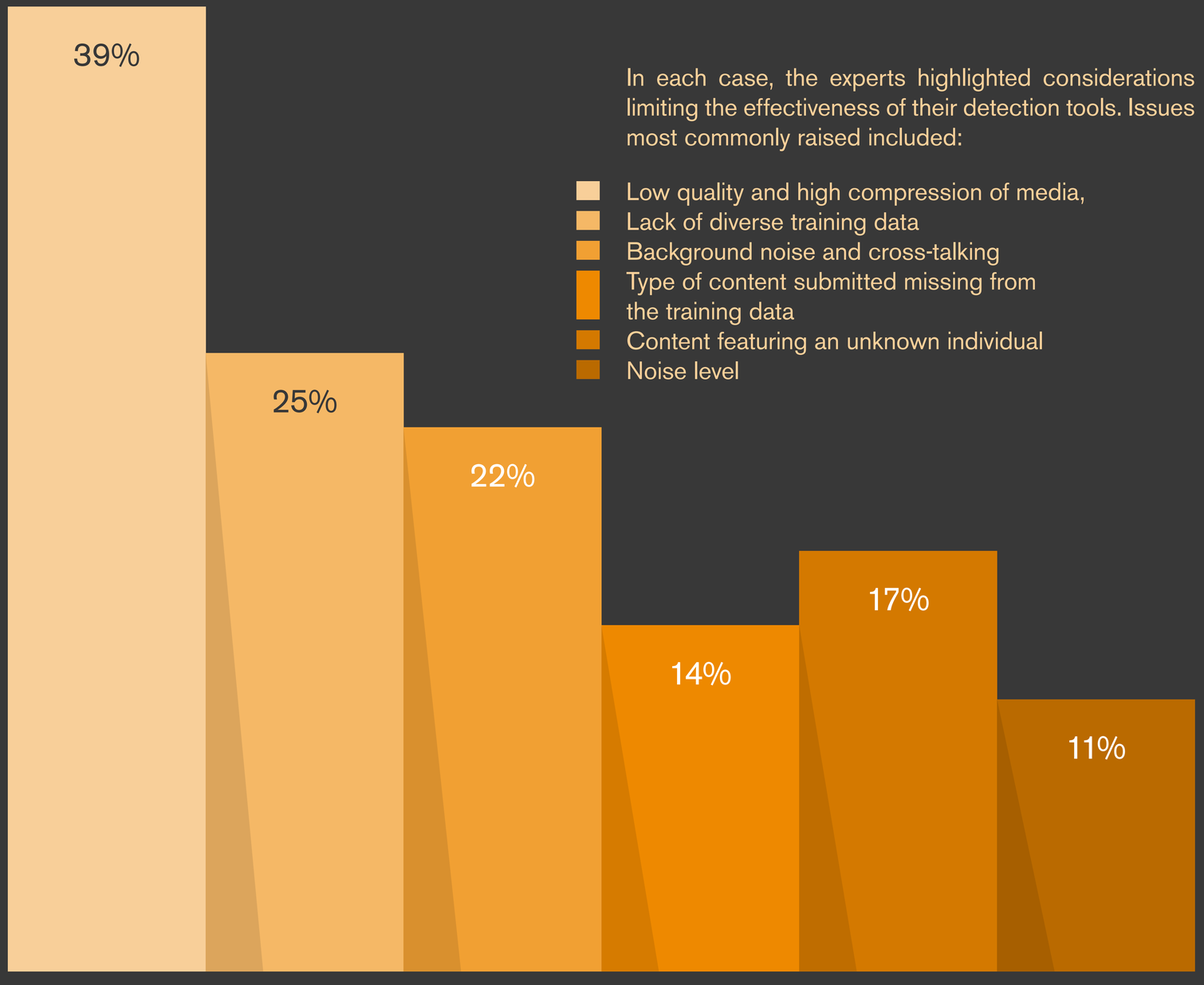
In 36% of cases analyzed, some of the most advanced detection tools failed to provide reliable results due to factors such as high media compression or a lack of representative training data, limiting the accuracy of detection.



2. Transparency and Explainability: An effective detection tool must offer clear, interpretable outputs that go beyond confidence scores and binary result labels. DRRF prioritizes responsible communication through tailoring technical insights from experts into accessible and transparent language for fact-checkers, which include information on the process and models' limitations. Such explanations later contribute to responsible reporting as some of the cases supported by the DRRF are shared with the public[17]. Detection tools should explicitly define their intended audience, objectives, their capabilities, and limitations. Such top-level transparency strengthens public trust, supports responsible AI literacy, and empowers journalists and fact-checkers to confidently interpret and communicate results. Explainability is not optional–it is central to the ethical use of detection tools.

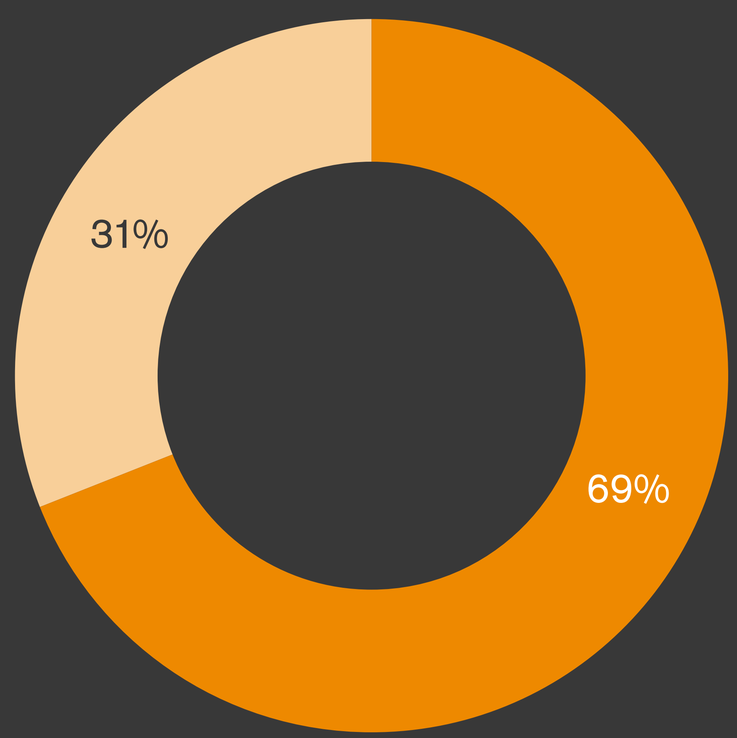
3. Targeted Accessibility and Usability: In Mexico, a viral image allegedly showing President López Obrador with El Chapo sparked confusion[18]. Public AI detection tools gave conflicting, inconclusive results, worsening mistrust. Though advanced tools clarified the truth, such resources are rarely available to frontline journalists and civil society groups, highlighting the pressing need for an AI detection tool to be accessible to its intended users, including communities with limited resources. While a tool may function in theory, barriers such as language limitations, technical skill requirements, connectivity issues, data privacy concerns, and costs can hinder its usability. Developers must clearly define their target audience and ensure the tool meets their needs. If the tool is to be used by a wide audience, it should accommodate users with varying levels of technical expertise.

4. Fairness and Representation: Fairness in AI detection tools must be prioritized at all stages, with the emphasis on development and deployment of the tool, and include considerations of how different elements influence the distribution of benefits to diverse communities. Fairness in training data is foundational, as it directly impacts detection accuracy and equity. Lack of training data representative of diverse demographics, languages, and contexts will result in the tool providing less reliable outcomes for content featuring demographics missing from the training data. In a case involving a video of the former President of Ghana[19], the DRRF experts noted that the combination of compression and darker skin color may lead the detection tools to reach inaccurate results. Additionally, the data used for training should be sourced in a fair and responsible way, as unethical data collection creates an opportunity for embedding bias in the detection outcomes.



In each case, the experts highlighted considerations limiting the effectiveness of their detection tools. Issues most commonly raised included:

- Low quality and high compression of media,
- Lack of diverse training data
- Background noise and cross-talking
- Type of content submitted missing from the training data
- Content featuring an unknown individual
- Noise level



Cases where the outlined issues directly affected the results of the detection tools.

Cases where the outlined issues did not directly affect the results of the detection tools.

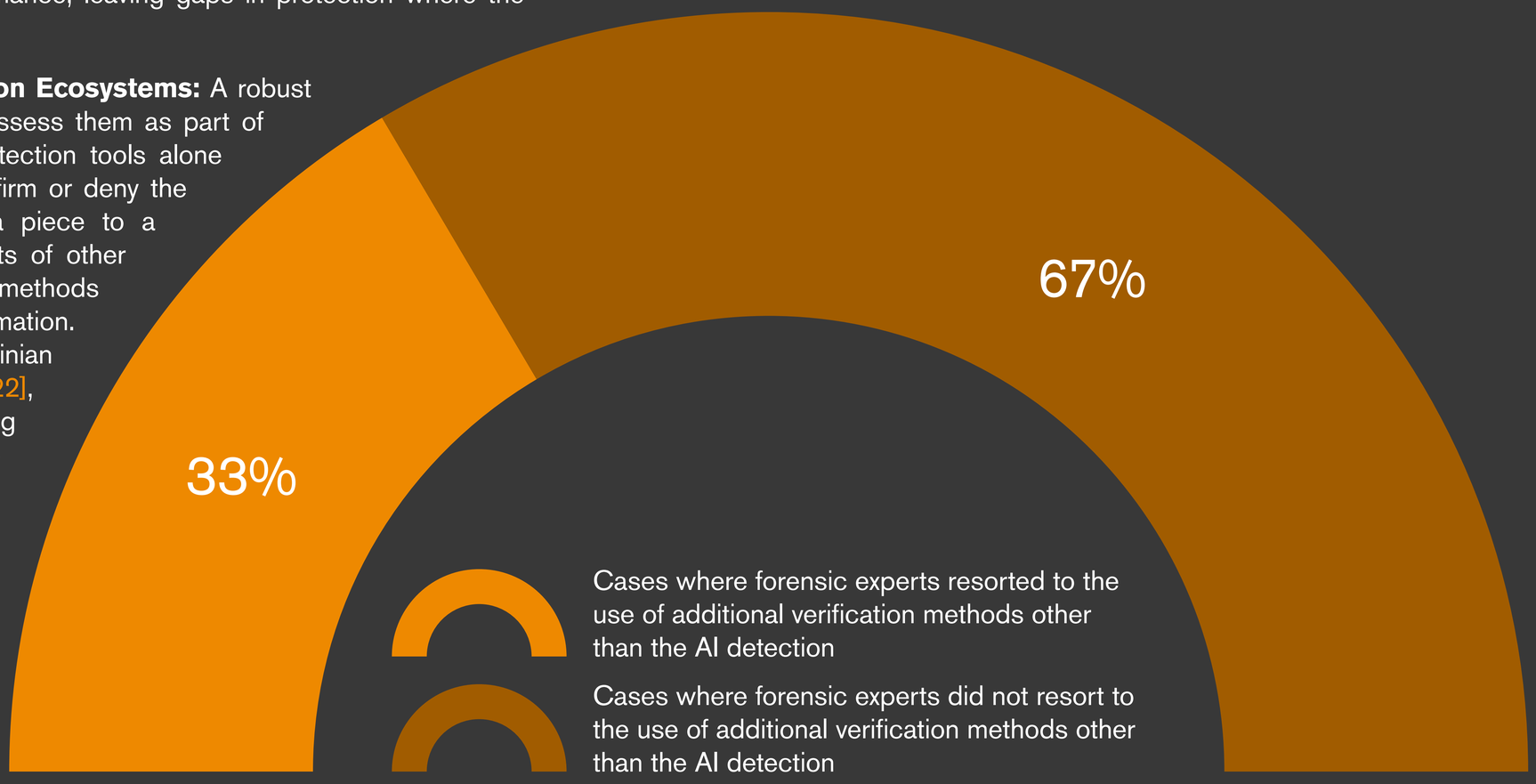
In 31% of cases, the outlined issues directly affected the results of the detection tools.

---

5. Durability and Resilience: As synthetic media evolves, AI detection tools must be regularly updated, evaluated and maintained to remain effective. DRRF experts inform the WITNESS team about updates and their impact on tool capabilities–for example, after receiving a number of cases with background audio, from Sudan[20], India[21] and Nigeria[16], some of the Force teams updated their models to handle noisy audio content. Ensuring that AI detection tools remain durable and resilient requires routine testing, adaptability to emerging technologies, and resilience against adversarial attacks. Consistent performance evaluations, such as reassessing previously verified cases and tracking accuracy over time, as well as continuous review of new classification tools and techniques, are critical. These efforts require substantial funding and institutional support. Without adequate resources, tools will degrade in performance, leaving gaps in protection where the stakes are highest.

6. Integration into Broader Verification Ecosystems: A robust evaluation of AI detection tools must assess them as part of an existing verification workflow. AI detection tools alone do not offer a definitive answer to confirm or deny the authenticity–rather they can provide a piece to a larger verification puzzle, which consists of other crucial techniques such as open-source methods and tracking relevant contextual information. When analysing an image of an Ukrainian soldier allegedly making a Nazi salute[22], the analysis teams received conflicting AI detection outcomes and resorted to manual verification methods, including leveraging additional image analysis tools, to reach a conclusive result. Detection tools contribute valuable signals, but their outputs must be weighed alongside other information sources, especially when results are uncertain.

In 33% of cases, forensic experts resorted to the use of additional verification methods other than the AI detection, including metadata analysis, human analysis of content and tracking related accounts.



Cases where forensic experts resorted to the use of additional verification methods other than the AI detection

Cases where forensic experts did not resort to the use of additional verification methods other than the AI detection

ACTIONABLE STEPS FOR DEVELOPERS AND POLICY ACTORS

Ensuring AI detection tools effectively serve the public interest requires concerted action:

For AI Detection Tools Developers

The AI developers are encouraged to evaluate AI detection tools against the key considerations stemming from each of the six core pillars of effectiveness. In particular, they are urged to:

1

Design AI detection solutions that respond to **real-world challenges** of deepfake detection, including low-quality, compressed, and heavily formatted content, and operate effectively across multiple languages, representations, and cultural contexts, addressing underrepresented demographics in training datasets.

2

Clearly communicate AI detection results, including expanding on confidence scores, dataset, and tool limitations, to foster credibility and functionality of the tool and ensure that it is accessible to users with varying technical expertise.

3

Establish internal policies for **regular updates, maintenance process, retirement protocols and durability benchmarks**, as well as regularly evaluate the tool performance across different demographics, contexts and against new developments in synthetic media.

4

Engage with a **diverse group of stakeholders**, including the AI regulators, standard bodies, and prospective global users, to support development of responsible AI regulations and standards, and ensure that AI detection tools align with principles of trustworthy and human-centric AI.

5

Conduct **inclusive and rigorous testing** with diverse teams representative of different stakeholder groups, including external red-teaming, to uncover global regional, contextual blind spots and vulnerabilities

6

Implement the TRIED Benchmark.
Scan or click on the QR code below to access the *TRIED Benchmark*

For International and Domestic Regulatory Bodies

1

Incorporate **sociotechnical considerations** into future regulations, codes of practice and other relevant legislation to ensure that detection tools evaluations reflect real-life applications and user experiences.

2

Design **accountability mechanisms** to safeguard fairness and accessibility considerations throughout the whole lifecycle of the detection system while accounting for security and safeguarding against adversarial attacks.

3

Encourage **global multistakeholder engagement** to ensure tools reflect diverse needs and contexts. Prioritize a collaborative and multistakeholder approach to the development of technologies and policies to ensure that the AI detection tools identify and respond to diverse needs and capacities.

For Standards Bodies

1

Set **minimum standards requirements** for detection tools to be considered truly effective from a sociotechnical perspective, based on the considerations outlined in this report.

2

Establish **guidelines on transparency, explainability, and fairness** to align detection tools with sociotechnical evaluations and global best practices.

3

Develop standards for **regular updates and durability benchmarks** for detection tools to address evolving AI technologies.

For Governments and Market Leaders

1

Push for the **implementation of detection standards** that ensure that AI detection tools deployed in public-facing contexts meet regulatory, safety and ethical standards.

2

Provide **secure and long-term funding mechanisms** that support research, development and long-term maintenance of detection solutions to ensure these can be equitably developed, regularly updated, and successfully adapted to promptly evolving synthetic media technologies.

3

Subsidize access to detection tools for key and targeted information and human rights actors.

4

Current and future funds should prioritize investment in **public interest AI funds** to advance technical solutions that promote a resilient global information ecosystem and facilitate the development of local AI resources, expertise and capacities, with a focus on underserved communities and high-risk information environments.

5

Invest in **training programs, workshops, and technical assistance** to ensure targeted stakeholders can effectively and responsibly leverage detection tools in real-world contexts.

CONCLUSION: LEADING THE WAY TOWARDS A GLOBAL RESILIENT INFORMATION ECOSYSTEM

Strengthening information integrity in the age of generative AI requires more than technical performence—it demands innovative, well-governed, and practically effective solutions. The *TRIED Benchmark* provides a critical foundation for evaluating and advancing AI detection tools that meet this moment. Policy actors and developers have a crucial role in driving this change through responsible development, informed regulations, robust standards, targeted funding, and fostering global cooperation. Urgent action is needed to ensure technology serves, rather than undermines, a resilient global information ecosystem.

With the support of

Digital Democracy Initiative

REFERENCES

1 WITNESS. https://www.witness.org/ (last accessed April 8 2025).

2 anlen s, and Wojciak Z. (2025) TRIED: Truly Innovative and Effective AI Detection Benchmark, developed by WITNESS. *preprint arXiv.* arXiv:2504.21489 [cs.CY].

3 anlen s. (2025) Human Rights Can Be the Spark of AI Innovation–Not Stifle It, Tech Policy Press. https://www.techpolicy.press/human-rights-can-be-the-spark-of-ai-innovation-not-stifle-it/ (last accessed April 8 2025).

4 WITNESS. Prepare, Don't Panic: Synthetic Media and Deepfakes. https://lab.witness.org/projects/synthetic-media-and-deep-fakes/ (last accessed April 8 2025).

5 Gregory S. (2021) Pre-Empting a Crisis: Deepfake Detection Skills + Global Access to Media Forensics Tools, *WITNESS Blog.* https://blog.witness.org/2021/07/deepfake-detection-skills-tools-access/ (last accessed April 7 2025).

6 Vazquez Llorente R, Castellanos J, and Agunwa N. (2023) Fortifying the Truth in the Age of Synthetic Media and Generative AI Perspectives from Africa, *WITNESS Blog* https://blog.witness.org/2023/05/generative-ai-africa/ (last accessed April 7 2025).

7 WITNESS. Deepfakes Rapid Response Force. https://www.gen-ai.witness.org/deepfakes-rapid-response-force/ (last accessed April 7 2025).

8 anlen s, and Vazquez Llorente R. (2024) Spotting the Deepfakes in this Year of Elections: How AI Detection Tools Work and Where They Fail, *Reuters Institute.* https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail (last accessed April 8 2025).

9 Frances-Wright I, Jacobs E, and Meyer E. (2024) Disconnected from reality: American voters grapple with AI and flawed OSINT strategies, *Institute for Strategic Dialogue.* https://www.isdglobal.org/digital_dispatches/disconnected-from-reality-american-voters-grapple-with-ai-and-flawed-osint-strategies/ (last accessed April 8 2025).

10 UN OHCHR. (2023) Taxonomy of Human Rights Risks Connected to Generative AI. https://www.ohchr.org/en/documents/tools-and-resources/taxonomy-generative-ai-human-rights-harms-b-tech-gen-ai-project (last accessed April 8 2025).

11 European Union Artificial Intelligence Act. https://artificialintelligenceact.eu/article/1/. (last accessed 7 April 2025).

12 Trustworthy & Responsible AI Resource Center. 3 AI Risks and Trustworthiness, *National Institute of Standards and Technology.* https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/#:~:text=Deployment%20of%20AI%20systems%20which,AI%20risks%20and%20reduces%20trustworthiness (last accessed April 8 2025).

13 OECD.AI Policy Observatory. OECD AI Principles overview. https://oecd.ai/en/ai-principles (last accessed April 8 2025).

14 UN OHCHR. (2012) Guiding Principles on Business and Human Rights : Implementing the United Nations "Protect, Respect and Remedy" Framework. https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights (last accessed April 8 2025).

15 OECD. (2023) OECD Guidelines for Multinational Enterprises on Responsible Business Conduct. https://doi.org/10.1787/81f92357-en (last accessed April 8 2025).

16 @GazetteNGR. *X*. https://x.com/GazetteNGR/status/1642218315685699586/ 2023. (last accessed April 8 2025).

17 Christopher N. (2023) An Indian politician says scandalous audio clips are AI deepfakes. We had them tested, Rest of World. https://docs.google.com/document/d/1XwnXm1_rf4_fj1SVpZTiJLHZElebN-AfmhlUsZyihR4/edit?tab=t.0 (last accessed April 8 2025).

18 @zeltzinjuareze. *X*. https://x.com/zeltzinjuareze/status/1762878195265638428/ (last accessed April 8 2025).

19 UTV Ghana Online. *Youtube*. https://www.youtube.com/watch?v=K1KYhenW3oM&t=596s, 2024. (last accessed April 8 2025).

20 @dahrinoor2. *X*. https://x.com/dahrinoor2/status/1771155215414067603/ (last accessed April 7 2025).

21 @mini_razdan10. *Archived from X*. https://drive.google.com/file/d/1XVRz3vVikXX9ttMTLZlQ7JOeUYZJQn9k/view?usp=sharing/ (last accessed April 7 2025).

22 Goldin M. (2024) Posts misrepresent a photo of a Ukrainian soldier balancing on his prosthetic limbs, AP News. https://apnews.com/article/fact-check-ukrainian-soldier-photo-nazi-salute-863799988341?fbclid=IwAR0w4ci7kN4tp0RKyfbrj_xH7VrxJxvckoY7PeAQ67WMZslobZZ5OCpHHLU (last accessed April 8 2025).