TRIED: Truly Innovative and Effective AI Detection Benchmark, developed by WITNESS

shirin anlen
WITNESS
shirin@witness.org

Zuzanna Wojciak
WITNESS
zuzanna@witness.org

Abstract

The rapid rise of generative AI and the increasing ability to create deceptive synthetic media pose a significant threat to public trust and information integrity, particularly in resource-constrained regions of the Global Majority. In the face of this growth, AI detection tools remain an invaluable resource for frontline information actors, but continue to underperform in real-world scenarios due to challenges related to explainability, fairness, accessibility, and contextual relevance. Drawing on frontline experiences, deceptive AI cases, and global consultations, this paper highlights the limitations of current AI detection tools and outlines how they must evolve to become truly innovative and relevant by meeting diverse linguistic, cultural, and technological contexts. It also introduces the Truly Innovative and Effective AI Detection (TRIED) Benchmark, a new framework for evaluating detection tools based on their real-world impact and capacity for innovation to guide inclusive AI detection tools development as well as support policy efforts towards AI procedures and standards reflecting global frontline community input.

1 Introduction

Deceptive AI content continues to present a serious threat to society by undermining information integrity and public trust in the media, governments, and the work of civil society. While promising advancements in provenance standards and metadata-based verification exist, systemic issues in implementation and emerging challenges [5] of trust and digital divides leave universal feasibility of standards impossible. As such, AI detection tools remain indispensable for verifying content integrity, exposing manipulation, and restoring trust in authentic media. To provide this critical support, detection solutions must not only excel technically but also perform effectively in diverse, real-world contexts.

Detection tools for synthetic media can be highly effective when applied to well-structured, high-quality content. An analysis of a deepfake video featuring a former advisor to the President of Ukraine [12] can serve as an example. The video depicted the individual speaking directly to the camera, with clear audio and no background noise—conditions under which AI detection tools deliver highly accurate results. However, such scenarios are the exception rather than the rule. Most manipulated media online exists in far more complex and variable contexts, where challenges like poor-quality footage, layered manipulation, and linguistic diversity reduce the tools' effectiveness. Despite these challenges, the effectiveness of AI detection tools is often evaluated using technical metrics such as accuracy, speed, scalability, and generalization. While important, these metrics fail to address the complexities of applying detection tools in messy, real-world scenarios.

Through global community consultations [34] and initiatives supporting frontline fact-checkers and journalists in verifying potential deepfakes [15], WITNESS [72] has observed a persistent "detection equity gap" [22] which highlights the disparity between the technical capabilities of AI detection tools

Article presented at the Second IASEAI Conference (IASEAI'26).

and their practical utility in the contexts where they are most needed. It is particularly pronounced in the Global Majority, where resource constraints, issues of representation in tool development, linguistic diversity, and contextual manipulation trends exacerbate challenges. To address this disparity between theoretical capabilities and practical applications, WITNESS presents a sociotechnical account of the effectiveness of AI detection tools. Informed by conversations with frontline information actors, AI detection experts and real-world cases analysed by the Deepfakes Rapid Response Force[15], this paper offers a comprehensive analysis of performance of AI detection tools in diverse global contexts, focusing on factors such as accessibility, transparency and explainability, fairness, contextual applicability, and the role of AI detection within a verification landscape. This analysis is supplemented by the *WITNESS' Truly Innovative and Effective AI Detection (TRIED) Benchmark*, a detailed checklist designed to guide AI detection developers, and other stakeholders, in evaluating whether the detection tools align with best practices for effective detection [59].

The *TRIED Benchmark* promotes a resilient and public interest approach to assessing AI detection tools, going beyond technical specifications and focusing on their real-world functionality. By prioritizing sociotechnical insights, the framework equips developers and policy actors with criteria to design innovative tools that effectively tackle synthetic media in diverse, real-world environments and drive responsible, human-centric AI innovation [57]. Only by addressing all aspects of effectiveness can AI detection tools fulfill their purpose: empowering communities worldwide to combat the growing threat of deceptive AI and safeguard information integrity.

The ***WITNESS' TRIED Benchmark: A Checklist for Truly Innovative and Effective AI Detection*** can be accessed in the preprint version of this article.

2 Background

Since 2018, WITNESS has led 'Prepare, Don't Panic,' the first civil society initiative tackling synthetic media threats on the information ecosystem. This work highlighted reliable and transparent detection tools as a key element of resilient approach to tackling threats posed by synthetic media, deepfakes, and multimodal generative AI [73]. To address the lack of access to such tools by critical information actors, WITNESS [72] launched the Deepfakes Rapid Response Force, an initative which connects frontline fact-checkers, journalists and civil society actors with leading media forensics and deepfake detection experts with experience in synthetic audio, video, and image detection to provide an evidence-based analysis of content threatening democracy and human rights [15].

The work of the Deepfakes Rapid Response Force (DRRF) illustrates the tangible impact of detection efforts. Its analyses have been widely reported, including by Human Rights Watch [69, 70], the Rest of the World [8] and GHOne TV News [65]. These contributions are a testimony to how effective detection can help mitigate crises in real-time, foster media literacy, and enhance public understanding of synthetic media. Conversations with frontline journalists reveal critical challenges faced by actors in the Global Majority, including fragile media ecosystems, ineffective detection tools due to gaps in training data for local languages, accents, and public figures, and the prevalence of manipulation tactics, that limit the practical effectiveness of AI detection and response strategies. Compounding these issues is a widespread lack of AI media literacy, which hinders the ability to interpret and trust detection outputs, and the trend to discredit authentic media content by claiming it had been AI-generated [17], further undermining the trust in the information ecosystem.

Strictly technical evaluations fail to capture these sociotechnical considerations, leading to an incomplete assessment of AI detection tools. A sociotechnical perspective expands the understanding of metrics of effectiveness by emphasizing the practical applications of AI detection tools in diverse contexts. They cannot be evaluated in a vacuum—they must be contextualized within these real-world applications. WITNESS' work highlights the importance of grounding assessments in the lived experiences of those navigating high-stakes scenarios. To provide adequate support to its users, AI detection approaches must address the needs and technical capacity of relevant actors [22], such as fact-checkers, frontline journalists, and civil society more broadly. A tool's success should be measured not only by its technical performance but also by how well it aligns with the practical experiences and challenges of actors dealing with deceptive AI.

This paper translates lessons from real-life AI detection interventions into an innovative evaluation framework. It was developed through examination of cases analyzed by the Deepfakes Rapid Response Force between March 2023 and April 2025 and a series of interviews with both fact-

checkers supported by the Force and the teams of experts that are part of the Force. The interviews explored different dimensions of effectiveness of AI detection tools and were guided by themes identified through the case analysis. In parallel, we reviewed existing AI detection evaluation frameworks considering how they reflect sociotechnical considerations. The process was informed by WITNESS extensive research into synthetic media and detection equity.

3 Related Work

As AI advances, its societal impact has come under increasing scrutiny. A growing body of work evaluates AI through a sociotechnical lens, exploring interactions between technology and society. The need to bridge the gap between AI systems and practitioners has been long acknowledged and promoted through solutions such as model cards [38], datasheets for datasets [19] and data statements [4]. Similar notions have been discussed in the context of safety [71], risks posed by synthetic content [6], algorithmic fairness [60], and algorithmic detectability of generative AI [62]. In policy, sociotechnical considerations have shaped frameworks emphasizing trustworthiness. The European Union lists seven key requirements of a trustworthy AI system [45], including transparency, robustness and fairness. Similar commitments are reflected by the National Institute of Standards and Technology (NIST) [44], the Organisation for Economic Co-operation and Development (OECD) [43], and, most recently, Article 1 of the EU AI Act calling for trustworthy and human-centric AI [1].

These values extend to synthetic media and detection policies. Initiatives like the Partnership on AI (PAI) outlined essential insights and recommendations for the field following the Deepfakes Detection Challenge in 2020 [46], as well as highlighted the need for a responsible and transparent approach to detection [46] and provenance [47] and advocated for equitable access to AI detection solutions [62]. There has been considerable progress in transparency methods as a way to ensure the authenticity of media, through initiatives such as the Coalition for Content Provenance and Authenticity (C2PA) [13] and the incorporation of a sociotechnical lens in the C2PA's Harms Modelling analysis [14].

With regard to detection, prevailing approaches to evaluation of the AI detection tools often do not account for their performance on real-world examples of deepfakes. Deepfake-Eval-2024 Benchmark [7] and the work by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) and South Korea's Sungkyunkwan University [31] are recent efforts that highlight the shortcomings of AI detection models when applied in real-life scenarios and call for a more comprehensive approach to assessments considering different factors that may limit the effectiveness of the tools. However, while an important step towards more effective AI detection, these works are still primarily technical and do not include broader sociotechnical considerations, such as accessibility, fairness and explainability.

The *TRIED Benchmark*, grounded in established principles of trustworthy AI and informed by expert analyses of real-world cases of deceptive AI, addresses this existing research gap and embeds sociotechnical perspective in the evaluations of AI detection tools. This approach goes beyond abstract notions of trust, offering a practical methodology to evaluate AI detection tools against their effectiveness and impact in diverse contexts. By providing concrete guidance, this paper bridges the gap between ethical AI commitments and real-world implementation, empowering policy actors and AI developers to ensure responsible and trustworthy deployment of these technologies.

4 Redefining Innovative Effectiveness of AI Detection Tools in Practice

Incorporating sociotechnical considerations into the evaluation of AI detection tools requires asking critical questions about the objectives and characteristics of a reliable AI detection system. The questions should explore:

Desired outcomes: what specific goals is the tool designed to achieve? How well does it perform across diverse real-world scenarios? Is its effectiveness aligned with the needs and experiences of the intended audience?

Unintended consequences: what potential risks or harms might arise, particularly for marginalized communities? How effectively are the tool's limitations and intentions communicated to users?

Implementation guidelines: how do elements constituting trustworthy AI translate into practical considerations? How can developers incorporate notions such as fairness and transparency into the development and deployment of the AI detection tools? How to ensure that end users understand

and trust the outputs provided? How effectively does the tool integrate into existing skills and verification workflows? Does the tool complement the verification process without positioning itself as a definitive solution? Are results communicated in ways that add value to broader investigations?

Diversity in development: is the tool trained and tested on diverse datasets that reflect different contexts, characteristics, and media types? Does its performance remain balanced across diverse groups, or are its limitations in diversity clearly communicated? Does the evaluation process include diverse perspectives in testing and validation to ensure its effectiveness in real-world applications?

Distribution of benefits: are the tools equally useful to different audiences? Are they accessible and affordable to diverse users, considering barriers such as languages, diverse levels of technical skills, and access to computational knowledge and resources?

Proactive approach: what mechanisms are in place to identify and address the tool's limitations or failings revealed in the course of the tool's application with in-the-wild content? Is the tool designed with long-term effectiveness in mind? How is the tool future-proofed for continuous improvement of audiovisual AI and the development of other detection tools? How is transparency balanced with protecting proprietary technology, information security, or preventing bad actors from reverse-engineering the tool?

Accountability and governance: who is responsible for oversight and accountability in the tool's deployment, implementation, future updating, and retirement process? How are ethical concerns addressed during the tool's lifecycle? Are mechanisms in place to address the tool's misuse or abuse?

Based on these considerations, this paper outlines six core pillars for evaluating innovative effectiveness in AI detection tools**.** These core considerations highly influence each other and are interconnected. A comprehensive approach to sociotechnical evaluation of AI detection tools has to account for balancing tradeoffs among these considerations.

4.1 Design AI Detection Tools to Deal with Challenges Posed by Synthetic Media in Real-World Scenarios.

Effective AI detection tools must handle the complexities typical of audiovisual AI in real-world scenarios. Insights from the DRRF highlight these challenges. Tools tend to perform best on high-resolution and near-original materials, yet much of the content analyzed by fact-checkers, journalists and human rights actors contain dynamic environments, noisy backgrounds, and high compressions due to social media and messaging apps. These platforms may re-encode a file's metadata at the time of uploading. Due to this encoding, social media platforms inadvertently create versions [16] from user-submitted images and videos for widespread dissemination. These reprocessed files are often of lower quality and may include additional data introduced by the platform itself. This gets even more complicated when content is shared and moves quickly across platforms, further altering its properties. These changes significantly hinder AI media detection, making it one of the biggest challenges for experts analyzing digital content. Some DRRF experts did not engage with cases where the quality of content was too low to provide reliable results. For example, a suspected AI video from the Philippines purportedly showed the president snorting cocaine [42]. The video's poor quality rendered AI detection tools ineffective. Such a common feature should not be overlooked and the tool should offer solutions to address this challenge. Since low-quality, low-resolution, and high-compression settings can obscure deepfake artifacts, AI detection tools must not only clearly communicate their limitations but also incorporate strategies to mitigate these issues. For instance, providing detailed explanations of these constraints and integrating support for complementary fact-checking and verification processes is essential for ensuring a comprehensive and trustworthy analysis (more on these strategies in the following sections). These shortcomings were also highlighted by the NIST–improving the performance of tools on post-processed content or media corrupted by noise or formatting [6] can be seen as opportunities to further develop and improve the AI detection solutions to ensure they can provide adequate support to actors relying on them.

Background noise, music, or cross-talking in audio recordings further complicate the detection process and lead to inconclusive results. In an audio clip allegedly featuring the President of Russia [41], the DRRF experts noted that the presence of only one speaker and the lack of background noise increased the reliability of the detection results. In contrast, loud background music in an audio recording allegedly featuring the voice of the President of Georgia [50] obscured crucial artifacts, complicating analysis. Similarly, in a case involving a purported phone call between an Indian

politician and an influencer [37], overlapping voices hindered the detection process. The presence of multiple voices can hide signs of manipulation. Similarly, a recording including a mix of manipulated and authentic voices can also confuse the detector into misclassifying the recording as a whole as either manipulated or authentic. These scenarios illustrate the need for tools to perform reliably on noisy or complex audio recordings, or ideally offer mitigation strategies as mentioned above.

The effectiveness of detection tools depends in great part on whether training datasets represent the type of content they are expected to analyze. Researchers behind the Deepfake-Eval-2024 Benchmark [7] observed how both open-source and commercial detection models performed poorly on real-world examples of deepfakes. However, the performance of open-source models significantly improved once they had been finetuned on a dataset representative of diverse in-the-wild deepfakes. The quality of detection results directly depends on whether the kind of content uploaded to the tool was present in the training dataset. For instance, to detect manipulation in low-resolution images, such images need to be included in the training data. Datasets should encompass diverse file types, augmentations, social media and messaging apps formats, and use cases to improve detection outcomes.

The diversity of types of media submitted by the fact-checkers presents another challenge to the detection tools. A suspected radio conversation recording (walkie-talkie) from the Sudanese civil war [11] can serve as an example. The DRRF's experts described how detection models had failed to conclusively analyze the recording since radio conversations often appear at a different frequency than the data used for training, resulting in fundamentally different outcomes. As AI continues to evolve, forensic experts are likely to encounter novel forms of synthetic media. While no tool can fully anticipate future deceptive content, developers should prioritize adaptability, enabling tools to collaborate with users to handle unexpected cases. Some detection tools can adjust their performance based on user input. For example, a person-of-interest detection system analyzes uploaded content by comparing it with verified samples of a specific public figure [10]. Such a technique can adapt in real time to what person it will be analyzing and its accuracy may depend on the amount of data uploaded by the user. It demonstrates a possibility of creating detection tools adaptable to specific content and detection aims. However, it is equally important to communicate the tool's limitations to the user if the tool was not trained on, and as a result is not intended to work, on a certain type of data (discussed more in the following sections).

Another recurring challenge involves authentic content being falsely discredited as AI-generated. For instance, a video of Ghana's Vice Presidential Candidate giving a public speech [64] was dismissed as a deepfake by the speaker himself. Although DRRF experts confirmed the video's authenticity, providing evidence-based analysis to support this conclusion was challenging—it is inherently more difficult to prove lack of manipulation than to identify it. These examples expose the gap between the public needs and current detection capabilities, underscoring the importance of designing tools that respond effectively to real-world scenarios.

4.2 Promote Transparency and Explainability of the Detection Results in a Clear and Detailed Manner.

The effectiveness of an AI detection tool heavily depends on the clarity, details, and quality of the information it provides. Transparent and detailed communication empowers users to responsibly interpret detection results and fosters trust in the tool. This requires a user-centric approach, grounded in humility and an awareness of the tool's capabilities, limitations, and intended audience.

Detection tools should explicitly define their strengths—such as the types of content they can analyze—and their limitations, including language constraints or contextual challenges. If tasked with analyzing content outside of their training scope, tools should clearly communicate that the results may be unreliable. Explainability of the tool rests on the annotation of the training data set, which is a time-consuming and resource-intensive process, in particular in the context of unexpected cases as described above. Therefore, it is important to be mindful of the effort required to ensure transparent results.

Many tools present detection results as confidence scores or binary outputs [58], but as early research had already indicated, it is not enough for users to interpret and trust these outputs [52, 36]. Fact-checkers frequently report difficulty interpreting confidence scores, limiting their utility in verification processes. To improve usability, tools should provide detailed explanations of their processes, including the top level information of the types of datasets were used and what types of manipulation

the tool is optimized to detect, potential limitations of the tool, identifying specific moments and regions where manipulation occurred, providing insights from the metadata and sharing tool's latest updates. NIST's 'Reducing Risks Posed by Synthetic Content' report underscores the importance of supplementing detection results with relevant contextual information [6]. Tools should highlight specific manipulated areas within content or provide metadata-based insights. It is key to note that the information provided to support the detection result should not include information on the actors who created and shared the content to protect their right to privacy [26].

There is a recognized tension between transparency and security. Providing granular details on certain aspects of these models, such as the exact training datasets or detection methodologies, may expose them to adversarial attacks. Research indicates that detailed knowledge of a model's training data can enable adversaries to craft targeted inputs that exploit vulnerabilities, bypassing detection mechanisms, rendering detection tools less effective over time [62]. This concern has been widely discussed in security research [29, 53, 54], particularly in adversarial machine learning, where attackers refine synthetic media to bypass known detection mechanisms. To mitigate this risk, sharing high-level information about detection models, such as their design objectives or the types of media they are trained to detect, is recommended, rather than specific manipulation patterns or source code**.** It is vital that governments and regulators understand this inherent tension. Overly stringent requirements for granular transparency, without considering the security implications, could inadvertently weaken the very tools designed to combat misinformation. A nuanced approach that prioritizes high-level transparency and user understanding, while safeguarding against adversarial exploitation, is essential for the responsible deployment of AI detection technologies.

The need to provide the user with a comprehensive information on the tool's model level has been reflected in model cards [38] and datasheets for datasets [19]. Such records detail the background information on the AI model and its workings, which allows different stakeholders to better understand and develop applications. Similar factsheets could accompany the detection tools adding application level and outlining additional information, including the tool's intended use, limitations, and evaluations across different factors. Currently, no settled guidelines exist for what detection tools should communicate, creating inconsistencies. There is an urgent need for establishing standards, such as uniform adoption of factsheets based on the framework of model cards and datasheets, which would ensure explainability, prioritize transparency, and improve user trust. Detection tools must also adapt language that reflects the complexities of AI-generated content [33] and moves beyond simplistic binary labels like "real" or "fake". This nuanced approach is essential in today's rapidly evolving media landscape, where AI-generated content doesn't inherently equate to "fake" and where clear, detailed communication is vital for fostering trust and usability.

WITNESS' DRRF exemplifies this approach by tailoring technical insights from experts into accessible and transparent language for fact-checkers. These explanations emphasize processes and limitations that may affect results, particularly in challenging cases where multiple tools produce conflicting outcomes. For example, in cases from Nigeria [18], Uganda [55], and Venezuela [39], low-quality recordings led to inconclusive results, undermining the tools' reliability**.** Addressing such challenges requires transparency about how specific factors, such as poor quality or background noise, may have influenced outcomes.

Providing users with clear explanations and transparent breakdowns of the detection process not only builds trust but also enables information intermediaries to educate broader audiences about generative AI and deceptive media. This fosters responsible communication, bridges gaps in AI media literacy—especially in underrepresented regions—and supports informed public discussions.

4.3 Ensure that AI Detection Tools are Accessible to Their Target Users.

An effective AI detection tool should be accessible to its intended users [21], 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. For instance, a voice recognition tool requiring Python programming knowledge was deemed exclusionary by a fact-checker. Additionally, and in line with the explainability considerations above, detection outcomes must be communicated in a clear manner to make the tools

accessible to a non-technical audience. Interoperability is another crucial factor. Some detection tools only analyze content generated with specific technologies, forcing users to navigate a fragmented landscape of AI techniques. This not only limits effectiveness but also adds unnecessary burdens on users. Connectivity and privacy challenges further impact accessibility. Tools reliant on stable internet connections or those with high storage and processing requirements may exclude users in areas with unreliable connectivity. Offline functionality would significantly enhance usability in these contexts. Data privacy and security concerns also affect accessibility. Lack of clear and transparent policies on data retention, storage and disposal procedures may deter users from uploading content containing personal or potentially sensitive information.

Cost remains a significant barrier. Paywalls—whether per verification or via subscriptions—restrict access, especially for resource-constrained communities. Notably, paid tools do not always guarantee reliable results. Researchers also noted a trend where free or low-cost tools become more expensive once the service acquires a big user base. Such practice again limits the tool's accessibility and has particularly detrimental effects if the journalists, fact-checkers, and human rights actors had consistently relied on the tool in their workflow.

However, accessibility without a robust implementation can amplify the risks posed by deceptive AI. For example, cases when online deepfake detectors incorrectly flag authentic media as fake, such as a video of Phyo Min Thein [23], the former chief minister of Yangon, Myanmar. The DRRF also dealt with a case from Mexico [75] where journalists tested two images—an original photo and a deepfake—using several publicly available detectors. The tools produced conflicting results, with one even identifying the original photo as manipulated. In comparison, the DRRF's experts have access to state-of-the-art tools and obtained a conclusive result that the image was manipulated. Publicly available deepfake detectors can also easily be tricked into providing incorrect results [58]. Cropping or lowering the resolution of an AI-generated image caused a detector to falsely identify it as authentic after initially labeling it as fake. Similarly, screenshots of AI-generated images or recordings of AI-generated audio often confuse these tools, as such transformations strip critical information used for verification [58].

These examples highlight the urgent need for advanced, reliable detection tools tailored to the specific contexts and requirements of fact-checkers, journalists, human rights defenders, and civil society more broadly. Such tools must be designed to handle real-world challenges effectively, fostering trust and ensuring their utility across diverse use cases.

4.4 Embed Fairness Throughout the Process of Development and Distribution.

Fairness in AI detection tools must be prioritized at every stage—from development to deployment—and encompass the distribution of benefits to diverse communities. Fairness in training data is foundational, as it directly impacts detection accuracy and equity. Research analyzing fairness in deepfake detection indicates how the distribution of attributes, such as race or gender, in the training dataset affects the performance of the AI detection tools [35].

For example, DRRF encountered a video of a Ghanaian politician [48] where detection was hindered due to darker skin tones and high video compression. Having analyzed the datasets of popular deepfake detectors, experts also noted how the datasets largely consisted of Caucasian faces [28]. This led to disparities where content featuring other demographics was more likely to be labeled as fake. Lessons from early implementations of facial recognition technologies (FRT) underscore these risks [24]. Unbalanced training datasets led to FRT disproportionately misidentifying people of color, exposing them to discrimination and harmful consequences [32]. Similarly, biased AI detection tools can produce unreliable results, amplifying human rights risks and disproportionately affecting already vulnerable communities. However, it is key to note the surveillance and human rights-related risks stemming from creating extensive image databases [40], in particular considering the more and more prevalent use of facial recognition by the police [63] and their growing access to databases not designed for policing purposes [68]. To address these concerns, both fairness and privacy must be central considerations throughout the development of detection tools.

Another challenge is the lack of data representing linguistic diversity and local contexts. Many tools are trained on limited data, such as English and Spanish-language content or "talking head" videos, such as those found in datasets like FaceForensics++ [56]. This narrow focus limits their effectiveness in analyzing diverse languages, accents, and content types. For instance, in a Sudanese

case [27], DRRF experts shared that they had struggled to verify a conversation recording due to the tool's inability to process the local language. This creates a significant barrier for communities from the Global Majority operating in underrepresented languages, who may be reluctant to use them. Similarly, identity-based tools requiring reference samples are ineffective when public content of the individual is scarce or nonexistent, as seen in a case involving a local journalist from Venezuela [49].

The data used for training should be sourced in a fair and responsible way with the consideration of intellectual property, data protection and biometric data laws. When creating and using training datasets, developers should ensure compliance with local data protection regulations, particularly with regards to the lawful grounds for data processing and safeguards against data leaks of sensitive information. It is worth considering how the funding of the tool can affect the type of data used for the training. Fairness also requires rigorous and inclusive testing processes. Developers should involve external teams from varied cultural, regional, and linguistic backgrounds to identify vulnerabilities and address blind spots. Global red-teaming efforts can ensure tools perform reliably across different contexts. Marketing also demands a fair and humble approach—the tool's accuracy should not be exaggerated [3] and should be supported by explanation so that users can rely on them responsibly.

4.5 Invest in Detection Durability.

As synthetic media grows increasingly sophisticated, AI detection tools must be regularly updated and evaluated to remain effective. Durability in detection tools is not just about maintaining functionality but ensuring their relevance in an ever-evolving landscape. This requires routine testing, adaptability to emerging technologies, and resilience against adversarial attacks. Similar issues are noted in cybersecurity research highlighting the struggle of threat detection systems to adapt to new challenges [30]. 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. Developers should also anticipate advancements in AI and design flexible systems capable of integrating new techniques.

The creation and maintenance of AI detection tools should also contain an element of anticipation —understanding the generative AI field and the pace at which technologies advance should prompt them to consider how the field and tech will change in the following years. Developers should future-proof their tools to guarantee that their approaches remain relevant in the long term and provide reliable results to those who depend on them. This includes preparing for long-term exposure to adversarial attacks, which can undermine the tools' reliability and model's security. These efforts require substantial funding. Limited financial resources can restrict the frequency and scope of updates, impacting the long-term effectiveness of detection tools. Alternative funding models and regulatory support are essential to ensure that tools are maintained responsibly and provide reliable performance to those who depend on them.

It is equally important that the current status of the AI detection tool is communicated to the user - a tool that has not been updated for a certain period may fail to detect newer cases of synthetic media or provide inaccurate results. In the DRRF, experts inform the WITNESS team about updates and their impact on tool capabilities. However, some fact-checkers admitted they overlook the role of maintenance when evaluating tools. If the tool does not seem useful, they will exclude it from their verification process and will not return to use it. To prevent this, developers must actively notify users about updates and improvements. Tools should have a well-defined retirement protocol in place before their release. If a tool is no longer maintained, it should be removed from public access [66] to prevent the risk of generating false outcomes. Ensuring transparency about the tool's status and updates fosters trust and encourages continued use by fact-checkers and other stakeholders.

The introduction of durability-related standards should address some of these challenges. Clearly established timelines for maintenance and criteria for removing outdated tools can help ensure reliability. Collaborative benchmarks comparing different detection solutions can also support the development of resilient systems capable of adapting to advancements in generative AI.

4.6 Leverage and Support the Integration of Tools Within a Broader Verification Ecosystem.

A robust evaluation of AI detection tools must position them as part of a broader verification landscape, examining their complementary relationship with other methods. 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.

Use of other verification methods alongside the detection tools is particularly important in cases where the latter cannot provide confident results. Considering that they rarely work with high-resolution original content, journalists and human rights researchers often resort to other techniques, such as using satellite imagery analysis, chronolocation, witness testimonies or looking for connected footage to verify a piece of media [74]. Fact-checkers also emphasize that AI detection tools are just one step in their verification workflows, often supplemented by additional tools and approaches.

Similarly, in analyzed cases, DRRF experts have often recommended using other methods to corroborate deepfake detectors' results. For example, in one specific case [20] where various AI detection tools provided conflicting results, the teams resorted to manual verification methods such as tracking related accounts, identifying similar image styles, and leveraging additional image analysis tools to analyze the image in question and were able to assert that it was authentic. The availability of diverse tools is particularly important in cases where the information actors aim to prove authenticity of a piece of media that is falsely purported to be AI-generated. As discussed in the previous sections, it is far more challenging to prove absence of manipulation than its presence. In such cases, the ability to support AI detection analysis with other tools and techniques provides users with more ways to assert credibility of a piece of media.

The need for reliance on diverse verification approaches is further underscored by the public confusion around the capabilities of AI detection methods. Due to lack of general understanding of synthetic media and AI detection techniques, some users approach the tools believing that they can detect any type of manipulation, including whether a piece of media was manually edited. Additionally, the evidence of AI manipulation itself is sometimes seen as evidence of intention to deceive, even when it does not affect the content of the message. For example, the use of AI-generated voiceover in a video does not automatically determine that the content of the message is also inauthentic; instead, the user should refer to other sources and methods to verify it. In a similar manner, researchers at Google highlight the need for a dual approach to verification [51], including both an analysis of how a piece of media was created or edited, and examining the background information and the context in which a piece of content was shared to fully evaluate its trustworthiness.

Given the diverse needs of users, specialized tools tailored to specific audiences, such as human rights defenders and journalists, can significantly enhance verification efforts. DeFake.app [61] is an example of a detection tool created through a collaboration with journalists, and tailored to their specific needs. Other tools, such as Amnesty International's Youtube Data Viewer [25], designed to extract metadata from YouTube videos, further demonstrate the value of tools created with specific use cases in mind. Similarly, collaborative initiatives like the Shakti Collective [9] in India offer valuable insights into fact-checking workflows in complex, multilanguage information ecosystems, providing guidance for developing tools that address diverse use cases.

AI detection tools play a vital role in helping fact-checkers verify and communicate the authenticity of content to the public. Tools that provide clear, evidence-based analysis strengthen the credibility of the verification process and enable fact-checkers to confidently and safely share their findings. For example, DRRF analyses have supported reporting on alleged deepfakes by allowing fact-checkers to corroborate [2] results using multiple tools and methods [67]. Transparent and detailed outputs also contribute to media literacy [8], fostering informed discussions about deceptive AI and enhancing public trust.

The context in which AI detection tools are used significantly affects how their results are interpreted. Confidence scores may be deemed adequate or insufficient depending on the user's expertise and the specific scenario. As discussed, the individuals often provide the missing context when using the tool and interpret the results within a broader verification process. Therefore, human insight is necessary to spot potential false results. In DRRF, the scores provided by the team's tools are complemented by analysis of human experts. In a case concerning an audio recording of the Ghanaian President ahead of the last general elections [48], it allowed the team to spot a false positive result produced by their deepfake detection tool. This highlights the necessity of human-assisted detection [6], where analysis is enriched by human expertise. Evaluating tool effectiveness should include metrics such as the time required to obtain results, ease of use for human operators, and comparisons between human-assisted and automated-only detection workflows.

While AI detection tools are generally part of a larger verification ecosystem, certain scenarios demand their use as the primary method for determining authenticity. Cases such as fabricated phone calls or audio recordings, observed by DRRF, illustrate instances where traditional open source investigative techniques, like reverse searches, are ineffective. Personalized and manipulated content of this kind often leaves fact-checkers reliant on AI detection tools as their primary solution. However, recognizing these tools' limitations in specific disinformation contexts underscores the importance of ensuring they are accessible, reliable, and advanced enough to meet the challenges of such scenarios.

5 Actionable Steps for Developers and Policy Actors

Strengthening AI detection is essential to promoting a resilient and safe information ecosystem, and requires multistakeholder engagement, emphasizing the need for technology companies and policy actors to promote technical solutions that respect human rights and embrace responsible innovation.

The AI developers are encouraged to evaluate AI detection tools against the key considerations outlined in this paper and follow the *TRIED Benchmark*. They are urged to prioritize designing tools that address real-world challenges of deepfake detection while ensuring equitable performance across multiple languages, representations, and cultural contexts. Tools should communicate results transparently, expanding on confidence scores, datasets and tool's limitations. Resilient detection tools require clear policies for regular updates, maintenance process, retirement protocols and durability benchmarks, as well as continuous evaluation. Engagement with a diverse group of stakeholders, including regulators, standards bodies, and global stakeholders will ensure that that AI detection tools align with principles of trustworthy and human-centric AI.

International and domestic regulatory bodies should incorporate sociotechnical considerations into future regulations, codes of practice, and other relevant legislation to ensure detection tools are evaluated under real-world conditions. They are encouraged to design and implement clear accountability mechanisms that safeguard fairness, accessibility, and security throughout the entire lifecycle of detection system and prioritize collaborative and multistakeholder approach to development of technologies and policies. Standards bodies are urged to set minimum sociotechnical standards requirements for detection tools, including clear guidelines on transparency, explainability, and fairness. They should also develop standards for regular updates and durability benchmarks.

Governments and market leaders should implement AI detection standards that ensure that tools deployed in public-facing contexts meet regulatory, safety and ethical requirements. They are also encouraged to provide secure and long-term funding mechanisms, and prioritize investement in public interest AI funds to advance solutions that promote a resilient global information ecosystem and facilitate development of local AI expertise.

6 Conclusions

AI detection tools could play a pivotal role in mitigating the risks of deceptive AI—but only if developed with a sociotechnical evaluation framework. Addressing this challenge requires AI detection tools that are not only technically advanced but can also adequately support users in diverse, real-world scenarios and complement existing verification approaches. Through analyzing real-life application of AI detection tools in diverse contexts, this paper outlined a comprehensive view of effectiveness of AI detection that prioritizes fairness, transparency, and adaptability. In doing so, it effectively bridges the gap between strictly technical evaluations of AI detection systems and their practical applications in high-stake contexts, and offers an innovative approach to ensure detection tools support communities most impacted by synthetic media threats.

Ensuring that the AI detection tools are contextually relevant, accessible and fair requires multi-stakeholder commitment and infrastructural support. A holistic approach is needed to address the considerations raised in this paper, and both AI developers and policy actors have a key role in advancing detection solutions that foster public trust and strengthen the global information ecosystem. The accompanying *TRIED Benchmark* [59] is designed to guide AI developers, policy actors and other stakeholders in evaluating whether the AI detection tools are effective from a sociotechnical perspective and support the diverse needs of their users while promoting adaptability, transparency, accessibility, fairness, and durability, and strengthening verification workflows. By adopting this framework, stakeholders can ensure detection tools meet the diverse needs of a global audience.

Acknowledgments

We would like to thank all the journalists, fact-checkers and researchers who have referred cases to our Deepfake Rapid Response Force, as well as the over forty experts and members of the Force who continue to provide WITNESS and their partners with pro bono computational models and expertise.

We also wish to thank the following individuals, all of whom provided insightful comments at different stages of the development of the framework: Henry Ajder, (Latent Space Advisory), Rabiu Alhassan (FactSpace West Africa), Mahsa Alimardani (WITNESS), Tracy Aminu (FactSpace West Africa), Ana Bregvadze (GeoFacts), Edward Delp (Purdue College of Engineering), Brandon Epstein (Magnet Forensics), Sam Gregory (WITNESS), Natalie Gyenes (University of Toronto), Rijul Gupta (Deep Media), Gabi Ivens (Human Rights Watch), Sean McGregor (UL Research Institutes), Claire Leibowicz (PAI), Jose Lopez (Intel Labs), Michael Macedo Diniz (Federal Institute of São Paulo), Ryan Ofman (Deep Media), Symeon (Akis) Papadopoulos (CERTH-IT), Maiko Ratiani (Myth Detector), Thomas Roca (Microsoft), Anderson Rocha (University of Campinas), Nikos Sarris (CERTH-ITI), Saniat J. Sohrawardi (Rochester Institute of Technology), Y. Kelly Wu (Rochester Institute of Technology), Raquel Vazquez Llorente (Google), Luisa Verdoliva (University Federico II of Naples).

We are grateful to the participants of the 9th NYU Computational Disinformation Symposium and the PAI AI and Media Integrity Steering Committee for feedback on this project's development. We also want to thank the reviewers who provided comments but wished to remain anonymous.

References

[1] European Union Artificial Intelligence Act. `https://artificialintelligenceact.eu/article/1/`. Accessed: 7 April 2025.

[2] Samedi Aguirre. Audio atribuido a margarita gonzález sobre programas sociales en morelos tiene indicios de manipulación digital. *Animal Politico*. `https://www.animalpolitico.com/verificacion-de-hechos/desinformacion/audio-candidata-morena-margarita/`, 2024. Accessed: 8 April 2025.

[3] Michael Atleson. Keep your AI claims in check. *Internet Archive*. `https://web.archive.org/web/20250115031325/https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check/`, 2023. Accessed: 8 April 2025.

[4] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[5] Jacobo Castellanos. Tomorrow's great digital divide: Content with or without provenance. *WITNESS Blog*. `https://blog.witness.org/2025/03/tomorrows-great-digital-divide`, 2025. Accessed: 7 April 2025.

[6] Bilva Chandra, Jesse Dunietz, and Kathleen Roberts. Reducing risks posed by synthetic content an overview of technical approaches to digital content transparency. Technical report, National Institute of Standards and Technology, 2024.

[7] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *preprint arXiv*. `https://arxiv.org/abs/2503.02857/`, 2025. arXiv:2503.02857 [cs.CV].

[8] Nilesh Christopher. An Indian politician says scandalous audio clips are AI deepfakes. we had them tested. *Rest of World*. `https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/`, 2023. Accessed: 8 April 2025.

[9] Shakti: India Fact-Checking Collective. `https://projectshakti.in/`. Accessed: 7 April 2025.

[10] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. *arXiv*. https://arxiv.org/abs/2204.03083/, 2023. arXiv:2204.03083 [cs.CV].

[11] @dahrinoor2. *X*. https://x.com/dahrinoor2/status/1771155215414067603/, 2024. Accessed: 7 April 2025.

[12] Facebook. *Facebook*. https://www.facebook.com/100002167747651/videos/457953553957488/?vh=e&extid=MSG-UNK-UNK-UNK-COM_GK0T-GK1C/, 2024. Accessed: 8 April 2025.

[13] Coalition for Content Provenance and Authenticity. https://c2pa.org/. Accessed: 8 April 2025.

[14] Coalition for Content Provenance and Authenticity. C2PA Harms Modelling 1.4. https://partnershiponai.org/resource/glossary-for-synthetic-media-transparency-methods-part-1/. Accessed: 8 April 2025.

[15] Deepfakes Rapid Response Force. https://www.gen-ai.witness.org/deepfakes-rapid-response-force. Accessed: 7 April 2025.

[16] Magnet Forensics. Getting to the source: Understanding metadata removal on social media. *Magnet Forensics Blog*. https://www.magnetforensics.com/blog/getting-to-the-source-understanding-metadata-removal-on-social-media//, 2024. Accessed: 8 April 2025.

[17] Isabelle Frances-Wright, Ellen Jacobs, and Ella Meyer. 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/, 2024. Accessed: 8 April 2025.

[18] @GazetteNGR. *X*. https://x.com/GazetteNGR/status/1642218315685699586/, 2023. Accessed: 8 April 2025.

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv*. https://arxiv.org/abs/1803.09010/, 2021. arXiv:1803.09010 [cs.DB.

[20] Melissa Goldin. 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/, 2024. Accessed: 8 April 2025.

[21] Sam Gregory. What's needed in deepfakes detection? Insights from WITNESS' global preparedness work and the partnership on ai's steerco on media integrity work on the deepfake detection challenge. *WITNESS Blog*. https://blog.witness.org/2020/04/whats-needed-deepfakes-detection/, 2020. Accessed: 8 April 2025.

[22] Sam Gregory. 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, 2021. Accessed: 7 April 2025.

[23] Sam Gregory. The world needs deepfake experts to stem this chaos. *Wired*. https://www.wired.com/story/opinion-the-world-needs-deepfake-experts-to-stem-this-chaos/, 2021. Accessed: 8 April 2025.

[24] Patrick J. Grother, Mei L. Ngan, and Kayee K. Hanaoka. Face recognition vendor test part 3: Demographic effects, nist interagency. *Internal Report, National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.IR.8280/, 2019. Accessed: 8 April 2025.

[25] Amnesty International. Youtube data viewer. https://citizenevidence.org/2014/07/01/youtube-dataviewer/. Accessed: 8 April 2025.

[26] Gabi Ivens and Sam Gregory. Ticks or it didn't happen: Key dilemmas in building authenticity infrastructure for multimedia. *WITNESS Report*. https://lab.witness.org/ticks-or-it-didnt-happen/#references/, 2019. Accessed: 8 April 2025.

[27] @johnnygould. *X*. https://x.com/jonnygould/status/1768345232184148139/, 2024. Accessed: 8 April 2025.

[28] Yan Ju, Shu Hu, Shan Jia, George H. Chen, and Siwei Lyu. Improving fairness in deepfake detection. *preprint arXiv*. https://arxiv.org/abs/2306.16635/, 2023. arXiv:2306.16635 [cs.CV].

[29] Muskan Khan and Laiba Ghafoor. Adversarial machine learning in the context of network security: Challenges and solutions. *Journal of Computational Intelligence and Robotics*, 4(1):51–63, march 2024.

[30] Antonio Lara-Gutierrez, Carmen Fernandez-Gago, and Jose A. Onieva. A framework for drift detection and adaptation in ai-driven anomaly and threat detection systems. *International Journal of Information Security*, 24:199, sept 2025.

[31] Binh M. Le, Jiwon Kim, Simon S. Woo, Kristen Moore, Alsharif Abuadbba, and Shahroz Tariq. Sok: Systematization and benchmarking of deepfake detectors in a unified framework. *preprint arXiv*. https://arxiv.org/abs/2401.04364/, 2025. arXiv:2401.04364 [cs.CV].

[32] Algorithmic Justice League. What is facial recognition technology? https://www.ajl.org/facial-recognition-technology/. Accessed: 8 April 2025.

[33] Claire R. Leibowicz. Regulating reality: Exploring synthetic media through multistakeholder AI governance. *arXiv*. https://arxiv.org/abs/2502.04526/, 2025. arXiv:2502.04526 [cs.CY.

[34] Raquel Vazquez Llorente, Jacobo Castellanos, and Nkem Agunwa. 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/, 2023. Accessed: 7 April 2025.

[35] Siwei Lyu and Yan Ju. Deepfake detection improves when using algorithms that are more aware of demographic diversity. *NiemanLab*. https://www.niemanlab.org/2024/04/deepfake-detection-improves-when-using-algorithms-that-are-more-aware-of-demographic-diversity/, 2023. Accessed: 8 April 2025.

[36] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[37] @mini_razdan10. *Archived from X*. https://drive.google.com/file/d/1XVRz3vVikXX9ttMTLZlQ7JOeUYZJQn9k/view?usp=sharing/, 2024. Accessed: 7 April 2025.

[38] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, page 220–229. ACM, 2019.

[39] MPvenezolano. *Youtube*. https://www.youtube.com/watch?v=T8RE-8OfFp4, 2024. Accessed: 8 April 2025.

[40] Daragh Murray. Police use of retrospective facial recognition technology: A step change in surveillance capability necessitating an evolution of the human rights law framework. *The Modern Law Review*, 87(4):833–863, dec 2023.

[41] @mustafali2001. *Tik Tok*. https://www.tiktok.com/@muss.mil/video/7438149251449752850/, 2024. Accessed: 7 April 2025.

[42] Solid PH News. *Facebook*. https://www.facebook.com/solidphnews1/videos/501528382261724/, 2024. Accessed: 7 April 2025.

[43] OECD AI Policy Observatory. OECD AI Principles overview. https://oecd.ai/en/ai-principles/. Accessed: 8 April 2025.

[44] National Institute of Standards and Technology. AI Risks and Trustworthiness. https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/#:~:text=NIST%20has%20identified%20three%20major,statistical%2C%20and%20human%2Dcognitive/, 2019. Accessed: 8 April 2025.

[45] High-Level Expert Group on AI. Ethics guidelines for trustworthy ai. *European Commission*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai/, 2019. Accessed: 8 April 2025.

[46] Partnership on AI. The deepfake detection challenge: Insights and recommendations for AI and media integrity. *Partnership on AI Report*. https://partnershiponai.org/a-report-on-the-deepfake-detection-challenge//, 2020. Accessed: 8 April 2025.

[47] Partnership on AI. Building a glossary for synthetic media transparency methods part 1: Indirect disclosure. *Partnership on AI*. https://partnershiponai.org/resource/glossary-for-synthetic-media-transparency-methods-part-1/, 2023. Accessed: 8 April 2025.

[48] UTV Ghana Online. *Youtube*. https://www.youtube.com/watch?v=K1KYhenW3oM&t=596s, 2024. Accessed: 8 April 2024.

[49] @PedroKonductaz. *Archived from X*. https://drive.google.com/file/d/1sN9OYILspCUaMf83LroQqsuqgLW4GRhT/view?usp=sharing/, 2024. Accessed: 8 April 2025.

[50] @phenixsaqartvelo. *Archived from Tik Tok*. https://drive.google.com/file/d/1EaB__rClt_pQCjH8M1StmNejNspWvSaA/view?usp=sharing/, 2024. Accessed: 7 April 2025.

[51] Google Public Policy. Determining trustworthiness through provenance and context. *Google Policy Paper*. https://static.googleusercontent.com/media/publicpolicy.google/en//resources/determining_trustworthiness_en.pdf/, 2024. Accessed: 8 April 2025.

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. *arXiv*. https://arxiv.org/abs/1602.04938/, 2016. arXiv:1602.04938 [cs.LG.

[53] Maria Rigaki. Adversarial deep learning against intrusion detection classifiers. Master's thesis, Luleå University of Technology, 2017.

[54] Ihai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *arXiv*. https://arxiv.org/abs/2007.02407/, 2021. arXiv:2007.02407 [cs.LG].

[55] @rwomchechen. *X*. https://x.com/rwomchechen/status/1860001051367342123/, 2024. Accessed: 8 April 2025.

[56] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arxiv*. https://arxiv.org/abs/1901.08971/, 2019. arXiv:1901.08971 [cs.CV].

[57] shirin anlen. 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/, 2025. Accessed: 8 April 2025.

[58] shirin anlen and Raquel Vazquez Llorente. 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/, 2024. Accessed: 8 April 2025.

[59] shirin anlen and Zuzanna Wojciak. TRIED: Truly innovative and effective AI detection benchmark, developed by WITNESS. *arxiv*. https://arxiv.org/abs/2504.21489/, 2025. arXiv:2504.21489 [cs.CY].

[60] Nimisha Singh, Amita Kapoor, and Neha Soni. A sociotechnical perspective for explicit unfairness mitigation techniques for algorithm fairness. *International Journal of Information Management Data Insights*, 4(2):6073–6105, nov 2024.

[61] Saniat J. Sohrawardi, Sovantharith Seng, Akash Chintha, Bao Thai, Raymond Ptucha, Matthew Wright, and Andrea Hickerson. Defaking deepfakes: Understanding journalists' needs for deepfake detection. In *Proceedings of the USENIX Symposium on Usable Privacy and Security*, 2020.

[62] Jonathan Stray, Aviv Ovadya, Claire Leibowicz, and Sam Gregory. Governing access to synthetic media detection technology. *Tech Policy Press*. https://www.techpolicy.press/governing-access-to-synthetic-media-detection-technology/, 2021. Accessed: 8 April 2025.

[63] Big Brother Watch Team. Big brother watch responds to facial recognition powers in new crime bill. *Big Brother Watch*. https://bigbrotherwatch.org.uk/press-releases/big-brother-watch-responds-to-facial-recognition-powers-in-new-crime-bill/#:~:text=The%20Bill%20allows%20the%20Government,risk%20of%20misidentifications%20and%20injustice, 2025. Accessed: 8 April 2025.

[64] Dadzie TV. *Youtube*. https://www.youtube.com/watch?v=p6VnexYxwrU, 2024. Accessed: 8 April 2025.

[65] GHOne TV. https://www.youtube.com/watch?v=OrFXklQz6bQ&t=2702s. Accessed: 8 April 2025.

[66] Matthew Maybe (umm maybe). Ai-image-detector. *Hugging Face*. https://huggingface.co/umm-maybe/AI-image-detector, 2022. Accessed: 8 April 2025.

[67] Haseem uz Zaman. Outgoing US President Biden did not confess to helping orchestrate pakistan 'regime change'. *Soch Fact Check*. https://www.sochfactcheck.com/us-president-joe-biden-did-not-confess-to-regime-change-conspiracy-pakistan-army-imran-khan/, 2024. Accessed: 8 April 2025.

[68] Big Brother Watch. Big brother watch briefing on clause 21 of the criminal justice bill. https://bills.parliament.uk/publications/53817/documents/4320#:~:text=Clause%2021%20also%20allows%20for,powers%2C%20with%20limited%20parliamentary%20oversight.&text=policing%20needs/. Accessed: 8 April 2025.

[69] Human Rights Watch. Burkina faso: Video shows soldiers disemboweling. *Human Rights Watch*. https://www.hrw.org/news/2024/07/26/burkina-faso-video-shows-soldiers-disemboweling-body, 2024. Accessed: 8 April 2025.

[70] Human Rights Watch. Iran: Growing evidence of countrywide massacres. *Human Rights Watch*. https://www.hrw.org/news/2026/01/16/iran-growing-evidence-of-countrywide-massacres, 2026. Accessed: 6 February 2026.

[71] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative AI systems. *arxiv*. https://arxiv.org/abs/2310.11986/, 2023. arXiv:2310.11986 [cs.AI].

[72] WITNESS. https://www.witness.org/. Accessed: 7 April 2025.

[73] WITNESS. Prepare, don't panic: Synthetic media and deepfakes. https://lab.witness.org/projects/synthetic-media-and-deep-fakes. Accessed: 8 April 2025.

[74] WITNESS. A community-based approach to visual verification to fortify the truth. *WITNESS Guide*. https://library.witness.org/product/community-based-approaches-to-verification//, 2024. Accessed: 8 April 2025.

[75] @zeltzinjuareze. *X*. https://x.com/zeltzinjuareze/status/1762878195265638428/, 2024. Accessed: 8 April 2025.