Combating AI-Generated Mis/Disinformation in African Elections

June 2025

CIPESA

TABLE OF CONTENTS

Executive Summary/Overview 3

Strategic Objectives 5

Program Overview 6

1. Public Engagement and Literacy 6
2. Collaborative Ideation and Co-Creation 8
3. Contextual Case Documentation 9
4. Community-Centred Verification 10
5. Training and Capacity Building 14

General Summary of Results and Successes 15

Major Challenges, Constraints and Lessons Learned 17

Challenges and Constraints 17

Lessons Learned 19

1. Expose The AI 20
2. Capacity Building for Journalists 20
3. Community Fakes 21
4. Spot The Fakes 21
5. Ruto Lies 22

Executive Summary/Overview

THRAETS is a civic-tech pro-democracy organisation focused on protecting democratic processes across digitally fragile democracies. This report presents the progress and impact of THRAETS' implementation of the ADRF-funded project titled “Safeguarding African Elections – Mitigating the Risk of AI-Generated Mis/Disinformation to Preserve Democracy.” With financial support from the Africa Digital Rights Fund (ADRF), managed by CIPESA, this initiative has enabled THRAETS to expand its work across the Global South, focusing on proactive, community-centred strategies to counter the emerging risks of AI-generated disinformation in electoral contexts. The program’s overarching goal was to strengthen the resilience of African electoral environments against digitally sophisticated manipulation, particularly through the use of synthetic media and deepfakes. To do this, THRAETS adopted a multi-pronged strategy to create sustainable solutions, built around three core pillars: public awareness, civic-tech innovation, and community engagement. Under this grant, key activities included:

1. The development of an open-source AI tracking and knowledge hub to crowdsource and monitor AI-generated content related to elections.
2. The conduct of red teaming activities to refine investigative methodologies and enhance AI detection tools and tutorials for identifying synthetic and manipulated media.
3. and comprehensive training programs for journalists and civil society organisations to detect and counter AI-generated disinformation tactics.

These efforts bore our different programs, which include the interactive "Spot the Fakes" quiz designed to improve digital literacy; the Ideathon fostering co-creation of anti-disinformation solutions; the "Ruto Lies: A Digital Chronicle of Public Discontent," which systematically documents and exposes content about the Kenyan president in light of the Reject the Finance Bill Protests; and the "Community Fakes" crowdsourcing platform, empowering citizens to report and verify deepfakes.

Participants and stakeholders have had significant successes in raising public awareness and digital literacy. The Ideathon successfully generated innovative solutions, and the "Ruto Lies" chronicle has provided a critical public record. The "Community Fakes" platform has shown promising engagement in leveraging collective intelligence for rapid verification to combat misinformation and disinformation. Overall, the program has made substantial strides in strengthening community resilience against AI-generated falsehoods and deploying practical tools for countering them.

Despite these achievements, the program faced considerable challenges, notably the rapidly evolving nature of AI-gen technology and persistent digital literacy gaps across various demographics. Constraints included the nascent policy environment around deepfakes and

varying internet access. Key lessons learned underscore the critical importance of multi-stakeholder collaboration, the necessity for continuous innovation and adaptation, and the profound impact of community engagement in combating disinformation. These insights will inform future strategies to sustain and expand our efforts in safeguarding democratic integrity.

Strategic Objectives

The strategic objectives of the 'Safeguarding African Elections – Mitigating the Risk of AI-Generated Mis/Disinformation to Preserve Democracy' are:

1. **Enhance Public Awareness and Digital Literacy on AI-Generated Disinformation**

To equip communities in digitally fragile democracies with the knowledge and critical skills to identify and resist AI-generated mis/disinformation, through interactive tools, public campaigns, and educational resources.

2. **Strengthen Investigative and Verification Capacity of Civil Society and Media Actors**

Build the technical capacity of journalists, researchers, and civil society organisations to detect, analyse, and counter synthetic media and algorithmic manipulation in electoral and civic discourse.

3. **Develop and Deploy Civic-Tech Tools for Disinformation Detection and Monitoring**

Create open-source, scalable, and sustainable technology solutions, such as crowdsourcing platforms and AI tracking hubs, to empower communities to detect and document election-related disinformation in real time.

4. **Foster Citizen-Led, Community-Centred Responses to Digital Threats**

Promote participatory verification processes and grassroots innovation (e.g. via ideathons and open data initiatives) that strengthen local agency in combating disinformation campaigns.

5. **Generate Context-Specific Evidence and Case Documentation to Inform Advocacy**

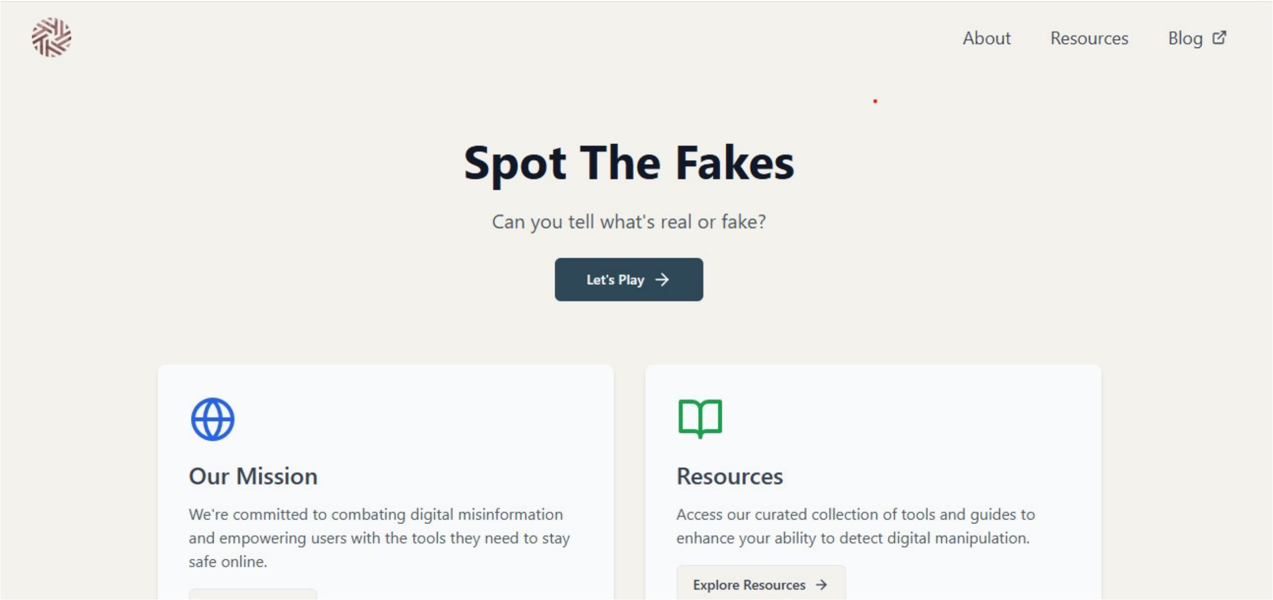
Conduct in-depth, data-driven case studies and research to document the real-world impacts of disinformation and AI-generated content on fragile democracies, informing advocacy and future intervention strategies.

Program Overview

Based on the strategic objectives above, we were able to create tools and programs whose key components included:

1. Public Engagement and Literacy:

Recognising that traditional capacity-building efforts, while impactful, often fall short of reaching broader audiences, THRAETS sought to engage the public through more accessible and interactive methods. To deepen understanding of the evolving ecosystem of AI-generated disinformation and misinformation, we launched *“Spot the Fakes”*, an engaging, gamified quiz designed to build critical media and digital literacy skills in a fun and accessible way. With different images that may or may not be AI-generated, Spot the Fakes trains the user to be able to distinguish between authentic and synthetic content, fostering greater awareness of the tactics and risks associated with AI-manipulated media. The quiz has become an effective entry point for broader public engagement, particularly among youth and casual digital media consumers, by simplifying complex issues into an interactive format.



Spot The Fakes

Can you tell what's real or fake?

Our Mission

We're committed to combating digital misinformation and empowering users with the tools they need to stay safe online.

Resources

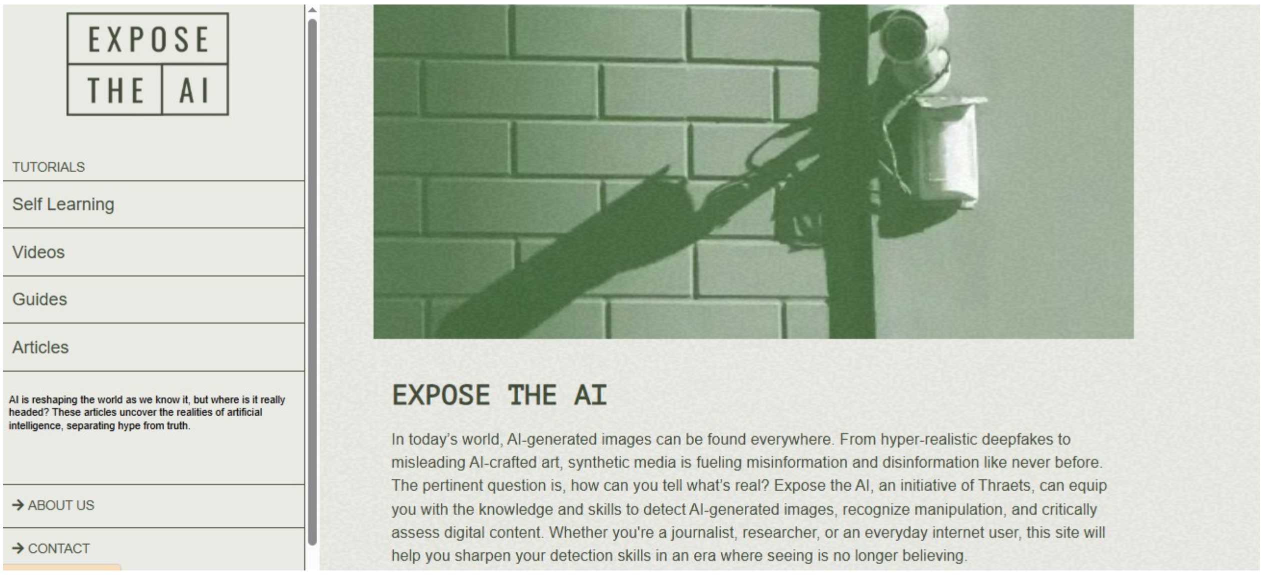
Access our curated collection of tools and guides to enhance your ability to detect digital manipulation.

The second program we launched under the component of Public Engagement and Literacy was “Expose the AI”, a website that promises to do just that. More and more, AI is becoming seamlessly woven into the fabric of our society. Generative AI introduces particularly troubling risks through its ability to create convincing deepfakes, which can erode public trust in visual evidence and enable targeted disinformation campaigns, some of which we have seen. The widespread availability of image manipulation tools threatens personal privacy and the security of one’s reputation, as anyone, from public figures to the everyday person, can be depicted in fabricated scenarios without their consent. Although video manipulation currently lags behind image synthesis in quality, rapid advancements indicate that we are approaching a future where it will be increasingly difficult for the average person to distinguish between authentic videos and AI-generated content.

Recognising the urgent need for transparency and public understanding, Threats developed Expose The AI. This website is a dedicated platform that is aimed at demystifying artificial intelligence, especially generative AI. This platform provides individuals with the tools and knowledge necessary to critically evaluate digital images, understand the risks associated with synthetic media, and protect themselves against deception in a constantly evolving information landscape.

With this platform, visitors will be able to discover a diverse collection of valuable, no-cost resources designed to empower individuals in recognising and comprehending synthetic content:

- Step-by-step tutorials on how to spot AI-generated images.
- Real-world case studies that show how synthetic media is used and misused.
- In-depth explainers on the ethical and societal implications of AI-generated content.



AI is reshaping the world as we know it, but where is it really headed? These articles uncover the realities of artificial intelligence, separating hype from truth.

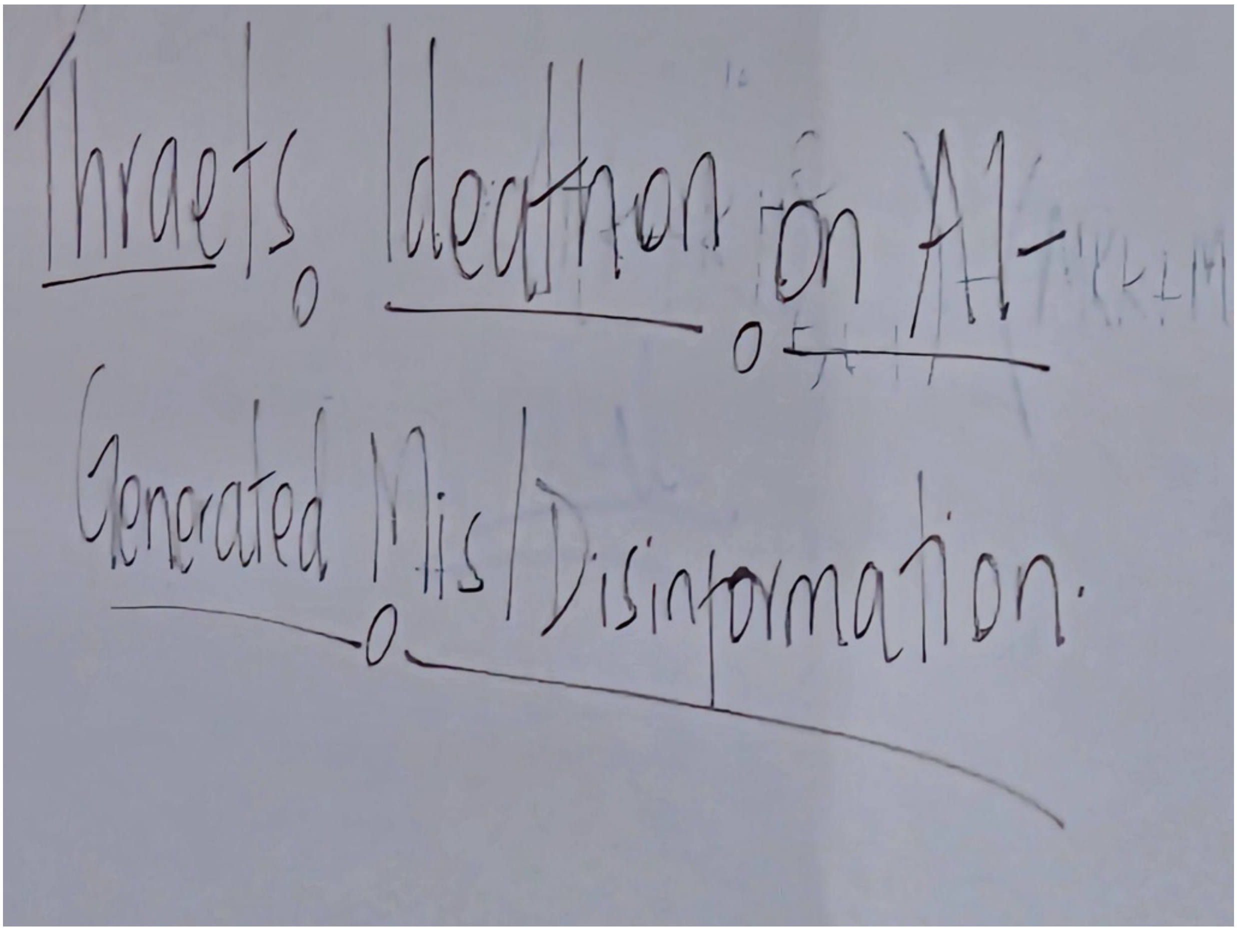
EXPOSE THE AI

In today’s world, AI-generated images can be found everywhere. From hyper-realistic deepfakes to misleading AI-crafted art, synthetic media is fueling misinformation and disinformation like never before. The pertinent question is, how can you tell what’s real? Expose the AI, an initiative of Thraets, can equip you with the knowledge and skills to detect AI-generated images, recognize manipulation, and critically assess digital content. Whether you're a journalist, researcher, or an everyday internet user, this site will help you sharpen your detection skills in an era where seeing is no longer believing.

2. Collaborative Ideation and Co-Creation

In this reporting period, we also executed a regional Ideathon, which brought together technologists, creatives, and activists to design new tools and interventions for combating disinformation, because AI-generated misinformation and disinformation pose a significant threat to democracy. They are spread by local/foreign state actors, businesses, and malign non-state actors bent on undermining and damaging free and liberty-loving democracies. Thraets, Outbox Hub and CIPESA held the AI Mis/Disinformation Ideathon

session to address the problem and find new solutions on the 20th of July, 2024. We encouraged teams to form to attack misinformation and disinformation from four tracks: government, business/technology, nonprofit, and Foreign Actor.



https://photos.app.goo.gl/qRLNVieFysxTgvgc8

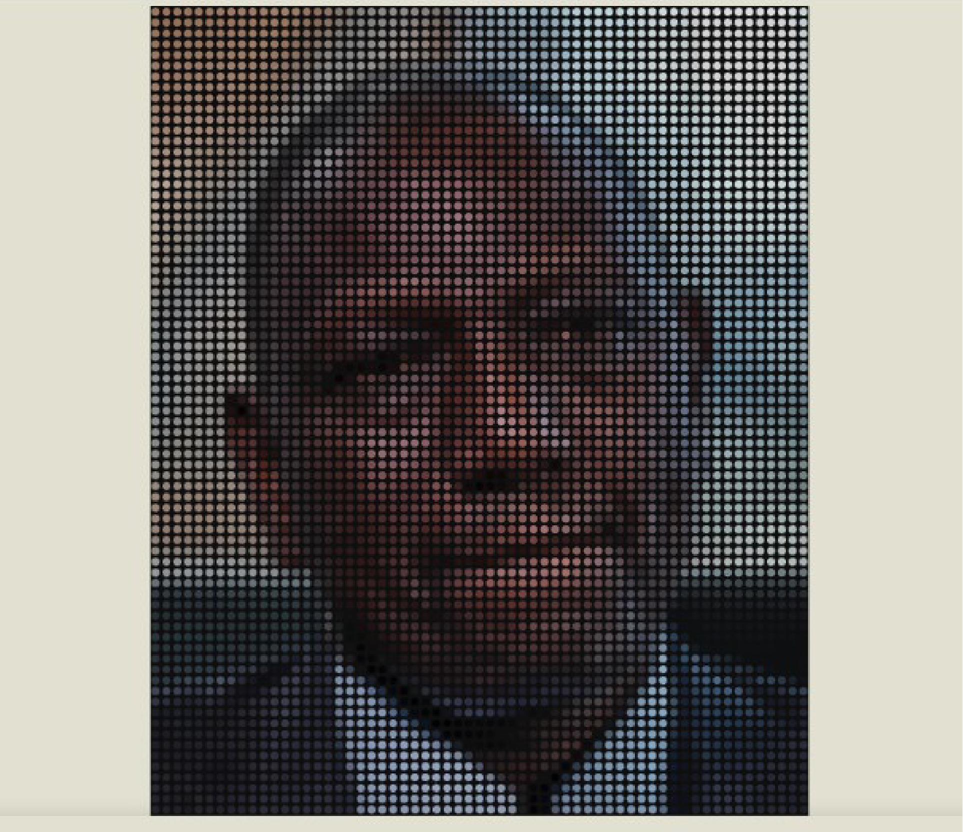
3. Contextual Case Documentation

In response to the Reject the Finance Bill protests in Kenya, THRAETS created the Ruto Lies: Portrait, which is a digital chronicle of public discontent. This investigative media archive examined the sentiments of the Kenyan populace on X (formerly Twitter) towards President Ruto, surrounding the reject the finance bill protests of 2024. In the intersection of Art and Politics, 'Ruto Lies' Portrait, an interactive webpage designed to highlight the grievances of Kenyan citizens against President William Samoei Ruto. Our research team meticulously analysed and filtered over 5,000 tweets to create this portrait. These tweets, collected during the #RejectTheFinanceBill demonstrations, contain various forms of the keywords 'Ruto' and 'lies'. This extensive dataset offers a rich resource for understanding public sentiment and the specific promises that are seen as unfulfilled. We released this research to

encourage developers and researchers to support the project by mapping these tweets to real claims or evidence of false promises made by President Ruto.

On the 17th of August 2024, Thraets participated in the HakiHack Hackathon, where we introduced and demoed the 'Ruto Lies' Portrait. The HakiHack Hackathon is a two-day event focused on developing tools to enhance democracy, good governance, and civic action. This event was a valuable opportunity for data scientists, statisticians, and civic tech researchers to engage with the Ruto Lies Portrait Project, access the data, and contribute to the mapping of tweets to real-world claims.

The *'Ruto Lies'* Portrait is both a visual spectacle and a powerful statement. Clicking on each dot allows users to read individual tweets that point out specific instances where the president's promises were perceived as unfulfilled.



4. Community-Centred Verification

In the Global South, political stability is often precarious, and elections can be influenced by mis/disinformation, which is much easier to access these days. The barrier to creating disinformation is no longer technical skill or cost; these tools are now readily accessible and often free. All it takes is malicious intent to create and amplify false content at scale. There is an increasing risk that these authoritarian regimes could weaponise AI-generated mis/disinformation to manipulate public opinion, undermine elections, or silence dissent. Through fabricated videos of political figures, false news reports, and manipulated media, such regimes exploit advanced technologies to sow confusion and mistrust among the electorate, further destabilising the already fragile democracies.

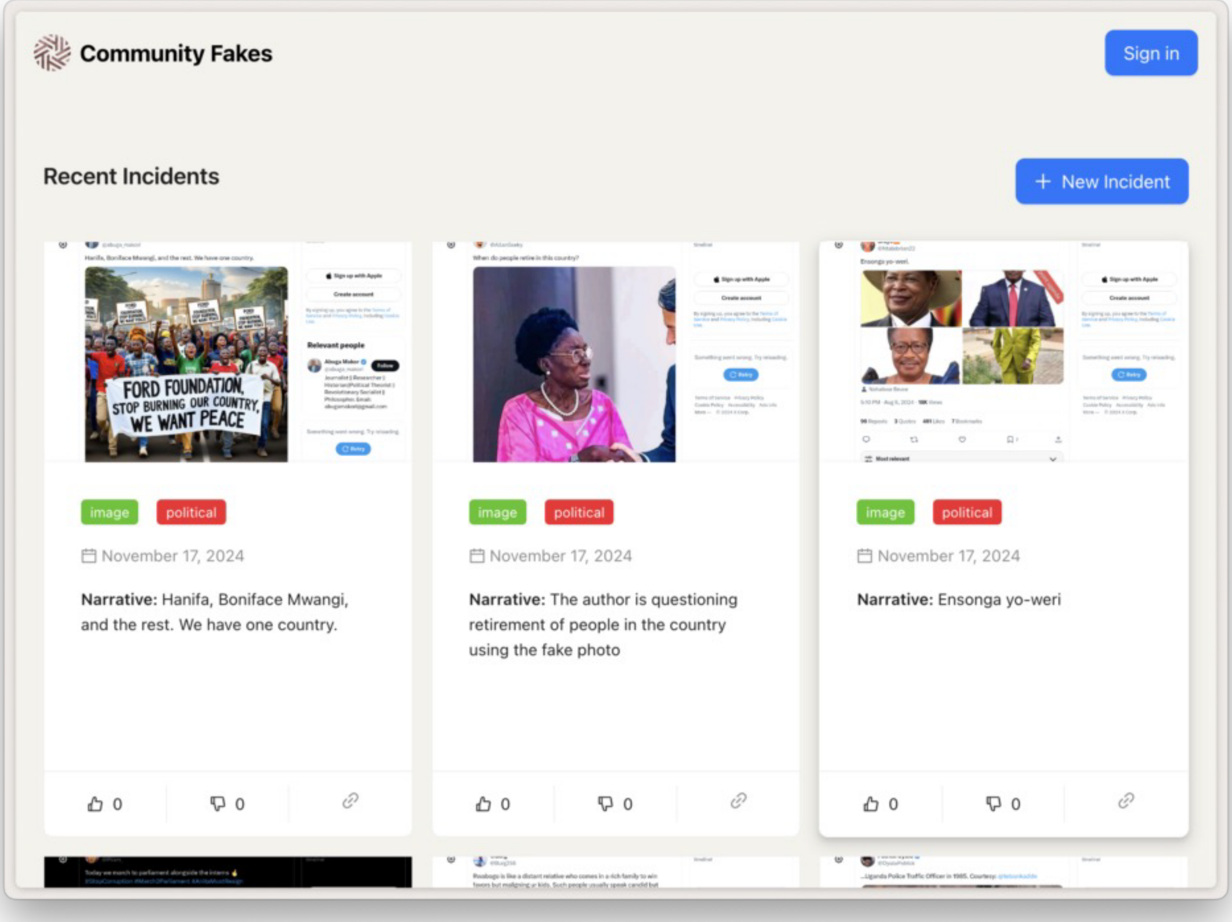
While social media platforms and AI companies continue to develop detection tools, these solutions remain limited in their ability to fully address the growing threat of synthetic disinformation, especially in culturally and linguistically diverse regions like the Global South. Detection algorithms typically depend on recognising patterns, such as unnatural blinking, mismatched lip movements, or anomalies in facial expressions, but these models are often trained on specific data from the West that doesn't account for nuances from the Global South. This limited scope enables deepfake creators to exploit local cultural cues and dialectical subtleties, producing media that automated detection systems struggle to accurately detect. This gap then leaves many communities vulnerable to disinformation, particularly during critical events like elections.

This rapid evolution of deepfake technology has shown the need for a stronger, combined approach that combines human and machine intelligence. Recognising this need, Thraets developed Community Fakes, an incident database and central repository for researchers. On this platform, individuals can join forces to contribute, submit and share deepfakes and other AI-altered media. Community Fakes amplifies the strengths of human observation alongside AI tools, creating a more adaptable and comprehensive defence against disinformation and strengthening the fight for truth in media by empowering users to upload and collaborate on suspect content.

Community Fakes crowdsources human intelligence to complement AI-based detection, and, in turn, this allows users to leverage their unique insights to spot inconsistencies in AI-generated media that machines may overlook while having conversations with other experts around the observed patterns. Users can submit suspected deepfakes on the platform, which the global community can then scrutinise, verify, and expose. This approach ensures that even the most convincing deepfakes can be exposed before they can do irreparable harm. Community Fakes will provide data sets that can be used to analyse AI content from the Global South by combining the efforts of grassroots activists, researchers, journalists and fact-checkers across different regions, languages, and cultures.

To further strengthen the fight against disinformation, Thraets also provided an API, allowing journalists, fact-checking organisations, and other platforms to programmatically access the Community Fakes database. We hope that this will streamline the process of verifying media during crucial moments like elections and enable real-time fact-checking of viral content. With the growing need for robust verification tools, this API offers an essential resource for newsrooms and digital platforms to protect the truth.

The launch of Community Fakes came at a critical time when the world is facing unprecedented challenges in combating disinformation and misinformation. Automated tools alone are not enough, especially in regions where AI may lack the necessary contextual understanding to flag manipulations. The combined power of AI and human intelligence offers the best chance to protect the integrity of information and safeguard democratic processes.


Community Fakes
Sign in
+ New Incident
Hanifa, Boniface Mwangi, and the rest. We have one country.
FORD FOUNDATION, STOP BURNING OUR COUNTRY, WE WANT PEACE
Sign up with Apple
Create account
Relevant people
Follow
Retry
When do people retire in this country?
Ensonga yo-weri.
Most relevant
Today we march to parliament alongside the interns
Uganda Police Traffic Officer in 1985.

Recent Incidents

image political

November 17, 2024

Narrative: Hanifa, Boniface Mwangi, and the rest. We have one country.

0 0

image political

November 17, 2024

Narrative: The author is questioning retirement of people in the country using the fake photo

0 0

image political

November 17, 2024

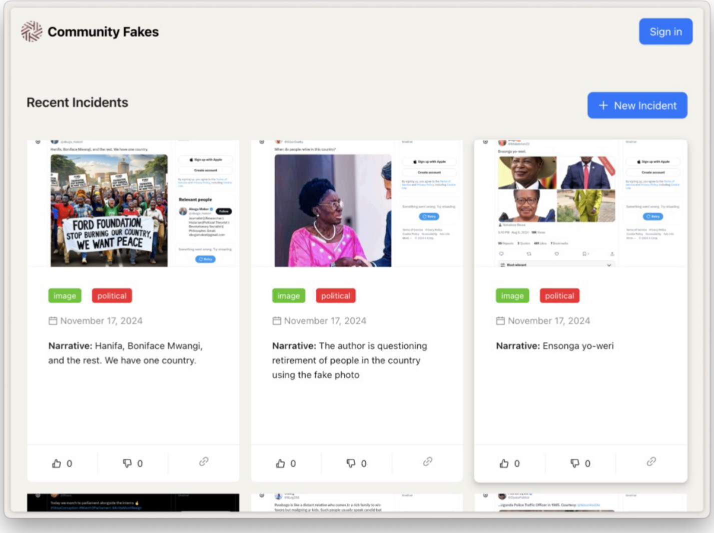
Narrative: Ensonga yo-weri

0 0

Using Community Fakes

1. Logging Into the Platform

Navigate to the login page of the PLATFORM website, community.thraets.org. You will be prompted to sign in using your email account. Select the desired account or add a new one if it's not listed. Once authenticated, you'll be redirected to the dashboard.



November 17, 2024

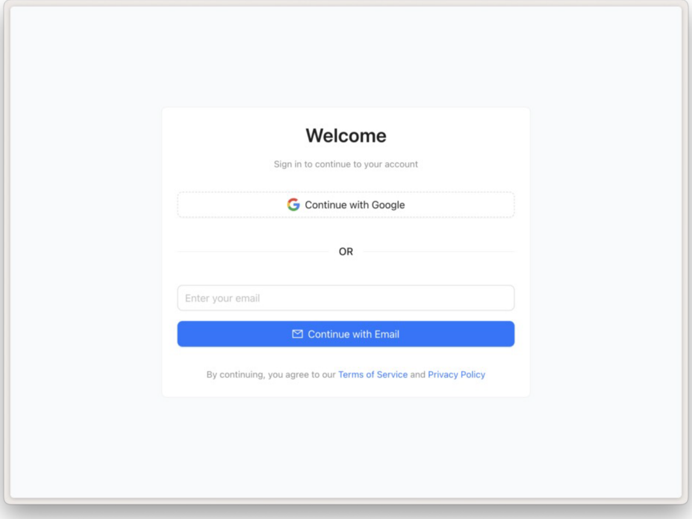
Narrative: Hanifa, Boniface Mwangi, and the rest. We have one country.

November 17, 2024

Narrative: The author is questioning retirement of people in the country using the fake photo

November 17, 2024

Narrative: Ensonga yo-weri

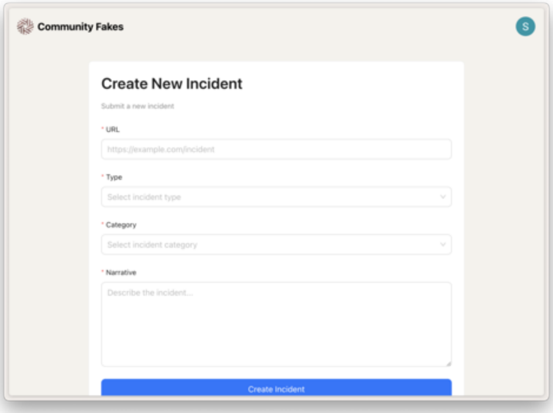


Sign in to continue to your account

By continuing, you agree to our Terms of Service and Privacy Policy

2. Editing an Incident

Edit the incident screen with fields like URL, Type, Category, and Narrative. Navigate to the “Edit Incident” page to update or create an entry. Fill in the necessary fields: Provide the link to the source of the fake or misinformation. Select the type of content (e.g., image, video, text). Categorise the incident appropriately (e.g., Political, Social, Health). Provide a clear and concise description of the issue. Toggle the “Verified” option to confirm authenticity and click Update Incident.



Community Fakes

Create New Incident

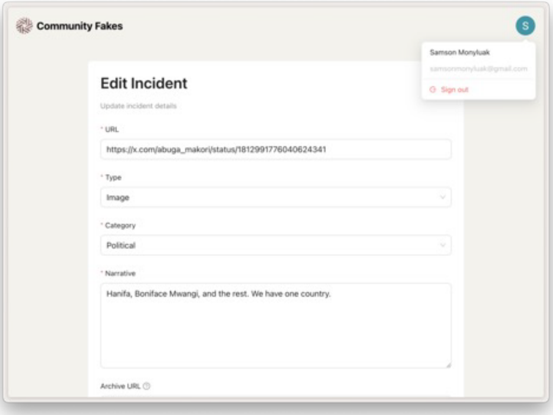
Submit a new incident

* URL: https://example.com/incident

* Type: Select incident type

* Category: Select incident category

* Narrative: Describe the incident...



Community Fakes

Edit Incident

Update incident details

* URL: https://x.com/abuga_makori/status/1812991776040624341

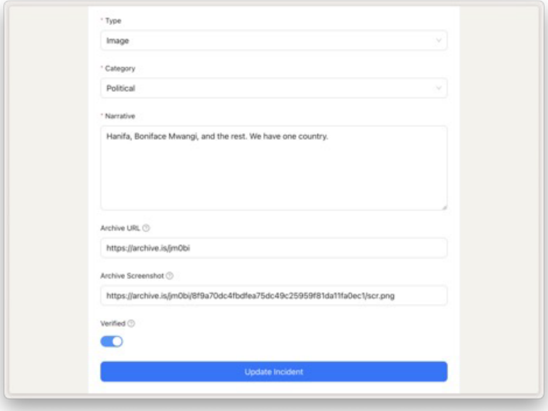
* Type: Image

* Category: Political

* Narrative: Hanifa, Boniface Mwangi, and the rest. We have one country.

Archive URL

Samson Monyluak
samsonmonyluak@gmail.com



* Type: Image

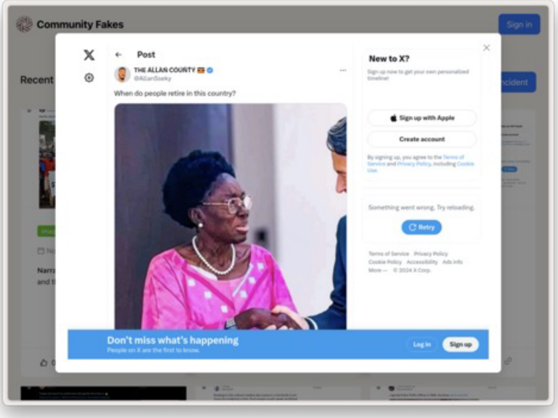
* Category: Political

* Narrative: Hanifa, Boniface Mwangi, and the rest. We have one country.

Archive URL: https://archive.is/jm0bi

Archive Screenshot: https://archive.is/jm0bi/8f9a70dc4fbdfea75dc49c25959f81da11fa0ec1/scr.png

Verified: ●



THE ALLAN COUNTY
@AllanSseky

When do people retire in this country?

New to X?

Sign up now to get your own personalized timeline!

By signing up, you agree to the Terms of Service and Privacy Policy, including Cookie Use.

Something went wrong. Try reloading.

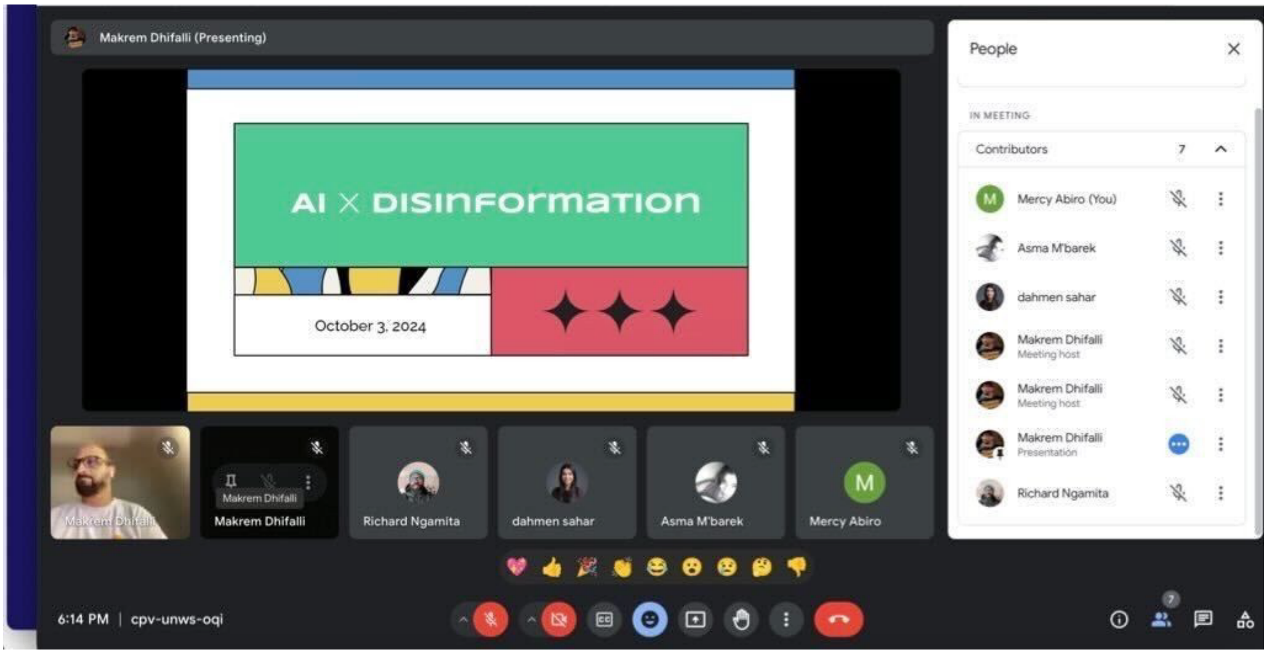
Don’t miss what’s happening
People on X are the first to know.

5. Training and Capacity Building

As artificial intelligence continues to influence how information is created and shared, the risks associated with AI-driven disinformation become increasingly significant. In light of this, enhancing journalists' skills is essential. In this reporting period, we were able to deliver practical workshops and tutorials for journalists and civil society actors to enhance investigative and verification capacities, especially in the context of fast-evolving AI technologies.

In August, we conducted an intensive training program for journalists as part of the Democracy Fellowship, a program funded by USAID and implemented by the African Institute for Investigative Journalism (AIIJ). This training aimed to enhance the journalists' capacity to leverage open-source intelligence (OSINT) tools in their reporting. Participants were equipped with the skills to access, analyse, and verify publicly available information from digital platforms, enabling them to uncover hidden narratives and expose critical issues with precision.

In October 2024, we partnered with eLab Research to conduct an intensive online training program for 10 Tunisian journalists ahead of their national elections. The sessions focused on equipping the participants with tools to identify and counter-tactics used to sway public opinion, such as detecting cheap fakes and deepfakes. Journalists were provided with hands-on experience through our engaging fake content identification quiz/game, available at *fakes.threats.org*. This training session provided journalists with the tools to identify and combat these threats, and this helped them prepare for election coverage, but also equipped them to protect democratic processes and maintain public trust in the long run.



AI × DISINFORMATION

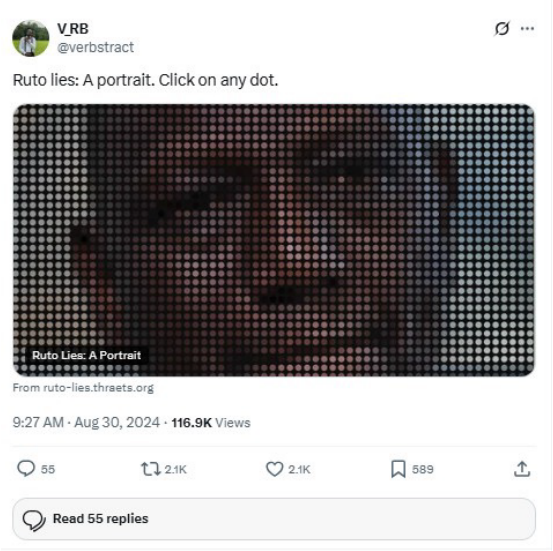
October 3, 2024

General Summary of Results and Successes

During the reporting period, THRAETS achieved significant success in raising awareness, building capacity, and developing tools to combat AI-generated disinformation in electoral contexts. The "Spot the Fakes" quiz is bound to become an effective entry point for digital literacy, particularly among youth. The "Expose the AI" platform has provided critical educational content and demystified complex AI technologies for a broader audience.

The Ideathon catalysed the co-creation of innovative, community-driven tools and fostered interdisciplinary collaboration across sectors.

Meanwhile, the "Ruto Lies" project not only documented broken political promises but also became a high-impact example of contextual disinformation tracking, drawing attention from media, civic actors, and the public. With over one hundred thousand views and interactions on X (formerly Twitter), we saw how much of an impact this particular project had on the netizens of Kenya, and this included researchers and journalists.



V_RB
@verbstract

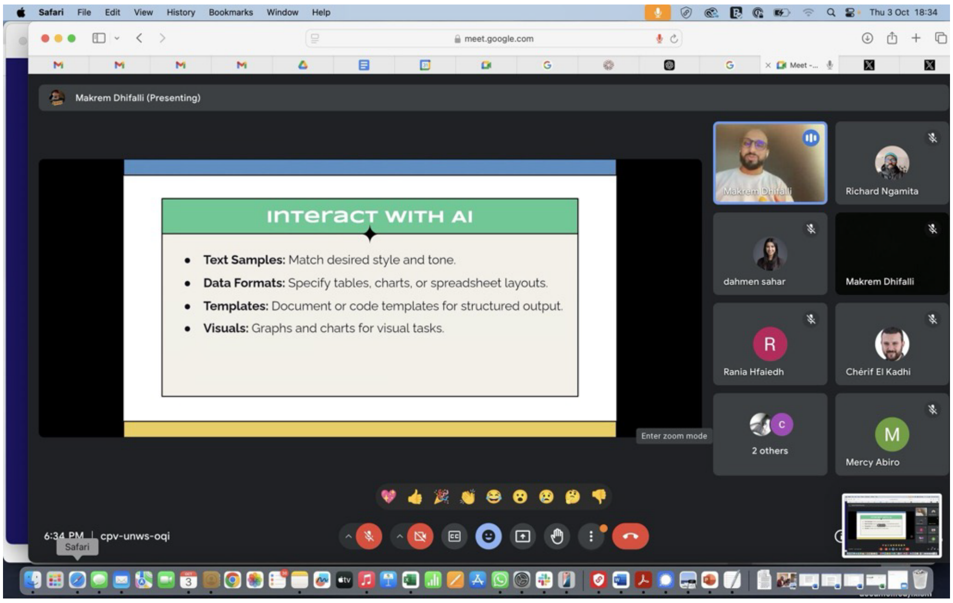
Ruto lies: A portrait. Click on any dot.

From ruto-lies.thraets.org

9:27 AM · Aug 30, 2024 · 116.9K Views

The Community Fakes platform demonstrated the power of combining crowdsourced verification with AI, empowering citizens to take part in safeguarding the information space.

Additionally, the training programs for journalists and civil society actors significantly enhanced local investigative and verification capabilities, with participants now better equipped to detect synthetic media and counter disinformation campaigns during elections.



INTERACT WITH AI

- **Text Samples:** Match desired style and tone.
- **Data Formats:** Specify tables, charts, or spreadsheet layouts.
- **Templates:** Document or code templates for structured output.
- **Visuals:** Graphs and charts for visual tasks.

To maximise outreach and engagement, THRAETS actively publicised the programs using multiple social media platforms, particularly LinkedIn and X (formerly Twitter). We also used a quarterly newsletter to share information about our programs. These platforms helped amplify our tools and findings, engage target audiences, and stimulate wider public discourse on the risks of AI in democratic processes.

Major Challenges, Constraints and Lessons Learned

Challenges and Constraints

1. Rapid Evolution of AI Technologies

The landscape of generative AI, particularly tools capable of producing hyper-realistic text, images, audio, and video, is evolving at a pace that far exceeds the capacity of most civil society actors to keep up. New technological advancements are released rapidly and often without transparency regarding training data, capabilities, or intended use cases. As a result, detection tools and countermeasures become outdated almost as quickly as they are developed. The increasing ease with which synthetic content can be generated makes it significantly harder for individuals to distinguish real from fake, further complicating efforts to build public resilience. In response, THRAETS developed and is continually updating educational content within our knowledge hub of EXPOSE THE AI. However, even these efforts face the risk of becoming obsolete in the near term due to the sheer speed of technological advancement. This volatility also challenged our knowledge hub, which required frequent revisions to address emerging AI-generated threat vectors and maintain relevance in a shifting disinformation ecosystem.

2. Digital Literacy Gaps Across Demographics

Africa's Internet penetration currently stands at just over 36 per cent, with only 473 million active internet users. Of these, 278 million users access the internet through their mobile phones. While digital literacy is growing in many parts of the Global South, large segments of the population still lack basic skills to critically assess digital content, especially AI-generated media. These gaps were particularly pronounced across generational lines (with older populations less likely to engage with interactive tools), between rural and urban communities, and among lower-income or less formally educated groups. Despite the development of engaging and accessible learning tools like Spot the Fakes, we have found that so many users lack the foundational knowledge to benefit fully from these solutions. This limited the broader societal impact of public awareness efforts.

3. Bandwidth and Access Constraints

In many parts of the Global South, users face unreliable internet connections, low bandwidth, and high data costs, particularly outside of major urban centres. These constraints made it particularly difficult for some of the target users, who would benefit most from these solutions, to access and engage with THRAETS' interactive tools, multimedia resources, or knowledge hubs. For example, the "Spot the Fakes" quiz and "Community Fakes" platform, though designed to be lightweight and mobile-friendly, still presented barriers for users with limited connectivity. This infrastructural challenge highlights the necessity for offline or hybrid solutions in future programming.

---

4. Sustaining Engagement and Verification Quality on the Community Fakes Platform

While the Community Fakes platform was developed as a novel crowdsourcing solution to identify and verify AI-generated deepfakes, it encountered specific challenges tied to both participation and data reliability. A key issue was sustaining active and consistent user engagement over time. Users were initially interested but did not return regularly to report or verify new content, limiting the platform's ability to build a robust, real-time repository of verified media. In addition, because the platform relies on open public submissions, ensuring the accuracy, relevance, and trustworthiness of flagged content posed ongoing difficulties. Without a centralised moderation team or a reputation system, there was occasional duplication, low-quality submissions, or subjective reporting that reduced the effectiveness of collective verification. Moreover, in fragile democracies where users may fear retaliation or surveillance, submitting suspicious content, especially politically sensitive media, requires strong trust in the platform's anonymity and safety protocols. These factors revealed that while community-based verification is powerful in theory, it requires complementary strategies such as incentive mechanisms, moderation support, privacy assurances, and continuous outreach to keep users engaged and confident in the system.

---

5. Digital Security Threats and Targeted Disruption

Following the release of two politically sensitive research, *Israeli Gas, Kenyan Tears*, which exposed the use of Israeli-supplied riot control agents during the 2024 #RejectTheFinanceBill protests, and "Ruto Lies: A Digital Chronicle of Public Discontent" THRAETS experienced a targeted Distributed Denial of Service (DDoS) attack on its main website. This attempt to overwhelm and take down our platform highlighted the increasing risks civic-tech organisations face when confronting powerful political or corporate interests through digital exposés. The DDoS attack temporarily limited public access to critical research and educational content, disrupted ongoing user engagement efforts, and forced an urgent reallocation of technical resources to restore and secure digital infrastructure. Beyond the immediate technical impact, the incident underscored a broader operational challenge: the need for resilient cybersecurity frameworks for civic-tech actors operating in fragile or hostile information environments. It also illuminated the importance of preparing for reactive censorship or information suppression attempts in the wake of high-impact content releases.

Lessons Learned

1. Agility Is Essential in AI Disinformation Work

The rapid evolution of generative AI requires programmatic agility and real-time adaptability. Static knowledge products, once considered foundational, are now insufficient in isolation. We learned that maintaining relevance in this space demands a strategy that is dynamic, frequent content updates and close monitoring of emerging AI tools. Investments in flexible infrastructure and open-source communities are key to sustaining these adaptive efforts.

2. Tools Must Be Paired with Foundational Literacy Campaigns

The effectiveness of public-facing tools such as Spot the Fakes and Community Fakes is limited by the wider digital literacy gaps in Africa. Awareness-raising campaigns and technical tools must be paired with fundamental education initiatives, particularly in under-connected or underserved communities. In the future, we should customise these efforts for different demographic groups, including youth, rural populations, and older citizens, to maximise their impact.

3. Infrastructure Realities Require Offline or Hybrid Options

Internet connectivity continues to be a significant barrier in the African context. Even lightweight tools can exclude some users in areas with limited bandwidth. Our experience has underscored the importance of designing interventions that include offline components, downloadable resources, and low-tech formats, such as SMS-based systems or printed guides, to help bridge the digital divide. Future programs should incorporate hybrid solutions to reach users beyond the urban, well-connected areas.

4. Community Engagement Must Be Sustained, Not Just Launched

Engagement on platforms like "Community Fakes" must be nurtured long after the launch. We learned that crowdsourcing alone is not self-sustaining—users need continuous touchpoints to remain involved. Trust-building, gamification, incentives, regular content updates, and moderation support are critical to ensure both platform engagement and verification quality. In fragile democracies, trust and safety mechanisms must also be explicitly communicated and built into the platform design.

5. Digital Security Is A Must

The DDoS attack we experienced highlighted the critical importance of having a strong digital security plan. Any organisation that operates at the intersection of technology and politics, particularly in fragile or repressive environments, must be prepared for targeted digital threats. Future projects should incorporate digital security as a fundamental part of platform architecture, staffing, and risk assessment protocols.

6. Political Sensitivity Requires Strategic Risk Mitigation

Publishing politically charged content, like “Ruto Lies” and “Israeli Gas, Kenyan Tears,” highlighted both the importance and the risks of contextual investigative work. We realised that operating in politically sensitive areas necessitates robust contingency planning, thorough legal reviews, effective crisis communication strategies, and partnerships with organisations that can offer protective support in the event of backlash.

7. Collaboration Enhances Impact Across Challenges

One consistent insight emerged from our experiences, and that is that collaboration enhances resilience and sustainability. Our strongest outcomes emerged from partnerships

with local communities, media organisations for literacy training, and digital rights coalitions for platform advocacy. Going forward, we will prioritise cross-sector coalitions to share the responsibilities of innovation, risk, and impact.

Capacity Gaps, Opportunities and Sustainability Plans

1. Expose The AI Capacity Gaps

- There is limited capacity to translate these emerging threats into local languages.
- There is a need for SEO Optimisation and performance scaling to withstand future coordinated attacks or traffic spikes.

Opportunities

- There is a high demand for contextualised AI literacy guides in local languages.
- There is also the potential to turn these guides into certified micro-courses for journalists, educators, students, and activists across the Global South.

Sustainability Plan:

- We need to develop a system for regular modular content updates that reflect the fast-changing AI landscape. For example, we could plan for quarterly guide updates and also video explainers for every new advancement in AI.
- We need to formalise partnerships with universities and other think tanks and digital rights organisations to share the burden of research, content generation, and multilingual adaptation.
- We could also create an editorial board to support crowdsourced, context-aware updates.

2. Capacity Building for Journalists Capacity Gaps

- Journalists on the African continent often lack the time or the technical baseline knowledge to fully grasp AI detection techniques.
- There is often limited post-training engagement to support the continued learning or applied practice of the OSINT techniques taught during the Capacity building.

Opportunities

- There is an opportunity to create regional cohorts for journalists tackling similar election disinformation risks. For example, we could create West Africa or East Africa-specific groups.
- Journalists trained through the THRAETS' capacity-building program could also develop case studies or alerts that could then be published through our platform.

Sustainability Plan

- For sustainability, a train-the-trainer model could work, building a core group of regional trainers who can lead workshops independently.

3. Community Fakes Capacity Gaps

- General public users often lack the technical knowledge and skills to verify deepfakes confidently.
- There is no dedicated team to review submissions and eliminate duplicates or low-quality entries.

Opportunities

- We could offer fellowships to digital activists or students to lead local Community Fakes deployments in their countries.

Sustainability Plan

- Introduce a reputation or rewards system to keep users consistently engaged (e.g., leaderboards, badges).
- Build a trusted network of moderators or “super verifiers” to maintain submission quality and vet disinformation claims.

4. Spot The Fakes Capacity Gaps

- Some users on older devices or slow networks still experience load issues, affecting usability.
- Current content may be too general for specific disinformation trends in particular countries or regions.

Opportunities

- Collaborate with mobile network providers or youth platforms to boost reach via sponsored data or campaigns.
- Launch customised versions of the quiz tied to national elections (e.g., “Spot the Fakes – Kenya Edition”).
- Create a mobile app version of the game for easy downloading by youth with smartphones.

Sustainability Plan

- Create downloadable versions of the quiz or SMS-based delivery to reach users with limited internet.
- We could partner with schools or digital literacy NGOs to integrate the quiz into civic education programs.
- Schedule regular “editions” of the quiz (e.g., every 3–6 months) to reflect current disinformation narratives.

5. Ruto Lies Capacity Gaps

- There is a need for more structured pipelines for mapping social media content to verifiable evidence.

Opportunities

- The archive is a powerful tool for journalists, scholars, and advocates pushing for transparency and political accountability.
- There is a potential to create a "Lies Chronicle Toolkit" to support other countries in documenting broken political promises using digital tools.

Sustainability Plan

- Turn “Ruto Lies” into a model for a series of digital archives documenting public sentiment and digital manipulation in elections.

Collaboration on International ICT Policy for East and Southern Africa (CIPESA)
+256 414 289 502
programmes@cipesa.org
@cipesaug facebook.com/cipesaug Linkedin/cipesa
www.cipesa.org