FemAI

CENTER FOR FEMINIST ARTIFICIAL INTELLIGENCE

WHITE PAPER



Table of Contents

Executive Summary **3**

Introduction **4**

Intended Audience For This White Paper **4**

Setting The Scene Of The "Super Election Year" 2024 ..... **4**

The Threat Of Deepfakes **6**

An Introduction To Deepfake Technology **6**

An Overview Of Current State Of Deepfakes **7**

AI And Elections: Lessons From South Korea **7**

A Perspective On Deepfakes Under The EU AI Act **8**

Why Are Deepfakes Problematic? **9**

Landscape Review Of Deepfake Detection Tools **11**

Guiding Principles For Best Practice **13**

Inclusivity 13

Robustness 14

Efficiency 14

Responsibility 15

Application Of The Principles **15**

Best In Class **18**

Inclusivity 18

Robustness 18

Efficiency 18

Responsibility 19

Conclusion **20**

Appendix **22**

This research was carried out on behalf of FemAI – Center for Feminist Artificial Intelligence.

Authors

Alexandra Wudel

Morgan Williams

Payal Padhy

All authors declare that they have no conflicts of interest.

The authors express their gratitude to Ali Gülerman for his insights and encouragement during the course of this research. We are also deeply appreciative of Nicolas Reitmeier whose continuous feedback has been instrumental in shaping this white paper.

The work stands on the shoulders of many dedicated individuals who have tirelessly contributed to this field, in a continuous effort to mitigate harms faced by digital citizens globally. Our endeavour would not have been possible without the commitment and dedication of the researchers and practitioners who have dedicated their careers to creating safe online spaces for everyone.

ATTRIBUTION: Wudel, A., Williams, M., & Padhy, P. (2024). ***Deepfake detection tools: An analysis for the super election year 2024.*** **FemAI - Center for Feminist Artificial Intelligence.**

Executive Summary

- This white paper aims to **bridge the gaps between AI policymaking, development, research, and civil society,** centering the needs of marginalized groups and providing insights for effective AI regulation, particularly for deepfake detection.

- FemAI uses an **intersectional feminist approach** to address power imbalances in AI, emphasizing the need for inclusive and tailored solutions for marginalized groups.

- The **'Super Election Year 2024'** is marked by significant elections around the globe, with deepfake technology posing renewed threats to democratic processes, particularly with regards to female and other marginalized election candidates.

- **Recent advancements in AI have simplified the creation of high-quality deepfakes,** leading to significant concerns over their use in political manipulation and the spread of disinformation.

- **Deepfakes disproportionately target women, particularly in pornographic contexts,** and pose severe threats to female politicians, deterring their participation in democracy.

- Various legislative measures, such as the EU's AI Act, Digital Services Act, and the UK's Online Safety Bill, reflect **global efforts to regulate deepfakes,** though challenges remain due to the rapid pace of technological advancements.

- Internationally, **we can learn from South Korea** and their proactive approach in regulating deepfakes during its recent elections. Their experience highlights the importance of public/private sector collaboration and targeted regulation to combat disinformation.

- **The white paper categorizes deepfake detection tools into four types**: Visual, Temporal, and Forensic Analysis; Physiological, Biometric, and Behavioral Analysis; Provenance and Integrity Verification; and Hybrid Approaches. We evaluated these based on four guiding principles for best practice.

- The **guiding principles are inclusivity, robustness, efficiency, and responsibility.** They are crucial for developing effective and ethical deepfake detection tools that protect marginalized communities and scale across societies around the world.

- **Call to action**: urgent and comprehensive action is needed, including global collaboration, AI literacy promotion, and the integration of deepfake detection tools into content moderation practices to protect democracies and marginalized groups from the threats posed by deepfakes.

Introduction

Intended Audience For This White Paper

FemAI aims to bridge the gaps between Artificial Intelligence (AI) policymaking, AI development, AI research, and civil society. This white paper provides insights that help build effective AI regulation through a collaborative approach between all stakeholders. It establishes the principles of best practice for deepfake detection tools and a path forward for the better protection of all individuals.

We aim to address existing power structures in AI, with a strong focus on marginalized groups. Our goal is to accelerate AI regulation to keep pace with technological developments and protect all individuals. We emphasize learning from and with each other.

Given the speed of technological change, regular reviews of tools are both possible and necessary. This paper is intended to guide the decision-makers in policymaking who are working on the threat of deepfakes and to motivate tech companies to engage in cross-sector collaboration.

FemAI uses an Intersectional Feminist approach. We use this method to avoid reductionism and address the discriminatory nature of AI in the best possible way. Finding a peaceful balance between organic nature and cutting-edge technological development is the main pillar of our approach, targeted to create a connection between our place of origin and the path that lies before us in the AI Age. A healthy curiosity for the future, coupled with never-ending respect for the past, guides our decisions. This is why this white paper also aims to raise awareness about the threat of deepfakes to democracies in our civil society generally.

This study analyzes the threats of deepfakes for female candidates, while not excluding all politicians, especially marginalized election candidates. To gather concrete insights, we applied statistics that are primarily available within the binary gender system. There are notable gaps in research beyond the binary gender system, alongside gaps in representation in training data, politics, and AI processes. However, when conducting our white paper, FemAI kept in mind that the findings should be applicable to all marginalized groups, without ignoring the fact that tailor-made solutions should be the ultimate goal.

Setting The Scene Of The “Super Election Year” 2024

2024 is shaping up as a pivotal year marked by significant global elections, where the intersection of AI and politics takes center stage. With over 60 countries electing leaders to govern nearly half the world's population, 2024 can be framed as a Super Election Year. Major economies such as the United States, United Kingdom, India, and the European Union countries will hold elections, with geopolitical tensions and rapid technological advancements framing the backdrop.

In 2016, the influence of data analytics and ad targeting on the US election raised global concerns. Cambridge Analytica, a data analytics firm, played a significant role in Donald Trump's 2016 presidential campaign by leveraging data from millions of Facebook users to create detailed voter profiles and target political advertisements. The use of personal data to manipulate voter behavior poses a threat to democratic processes. The scandal revealed how data-driven strategies could be used to influence elections, potentially undermining the integrity of electoral systems, not only in the U.S. but globally. At that point, relevant regulations were not translated into the digital world yet. However, the scandal around this event piled pressure on regulators to impose effective rules.

Since 2016, the world has evolved at a fast pace. AI has become a matter of global concern while also offering hope for the future. After entering the AI Age, elections in 2024 need to be considered within this new context.

Technological advancements, particularly in AI, have brought sophisticated tools such as deepfake technology further to the forefront, raising concerns over their use in disseminating disinformation. Deepfakes represent a modern twist on the much older problem of disinformation, posing renewed threats to the integrity of

democratic processes. We are already seeing in India's 2024 general election, the world's largest democracy, that deepfakes have been employed to woo voters and sway public opinion, demonstrating both their utility and ethical dilemmas.

This white paper will explore challenges posed through the availability and widespread use of AI in the context of elections. From their deployment in political contexts to their use in pornographic imagery, this analysis outlines the disproportionate impact on female election candidates without ignoring the threat to all election candidates, and democracy as a whole.

In examining these issues through an intersectional feminist lens, we express the need for comprehensive solutions that prioritize the stakeholders currently underserved within the AI ecosystem. The intersectional feminist approach applied is grounded in FemAI's past publication titled "Power Imbalances in Society and AI: On the Need to Expand the Feminist Approach."

At this critical moment, several key legislative measures have emerged. In the EU, the Digital Services Act (DSA), and the Artificial Intelligence Act (EU AI Act) as well as the Gender-based violence directive (GBVD) are focal points for deepfake regulation. The Online Safety Bill in the UK and the TikTok divestment directive (and several state-level AI regulations) in the US, confirm the need for effective, fast regulation. In South Korea, the legislature revised their Public Elections Law ahead of voting day to ban election-related deepfakes. These actions reflect a growing global-scale concern to regulate technology and mitigate its potential misuse. However, it has become evident that the slow nature of regulatory procedures is a challenge in the context of exponential AI developments and hinders their effectiveness in addressing the challenges posed in this super election year, 2024.

The EU's call on tech firms to outline their plans to tackle deepfakes ahead of their elections It underscores the awareness of governments about the threat from new technology while highlighting the urgency for them to take swift action in the absence of formal regulation and to address the issues proactively themselves.

This white paper aims to close the informational gaps between the relevant stakeholders in policy making, technology companies, and civil society. It does so by analyzing the spread of deepfake technology, spotlighting examples found in the recent South Korean elections, as well as centering the needs of marginalized groups when evaluating current deepfake detection tools as part of our landscape review. South Korea is one of the most affected nations in terms of deepfake manipulation, according to a recent study. South Korea's 2024 General Election, held on April 10, offers valuable lessons for the rest of the world. We will be summarizing key learnings from it and suggest ways to implement them in the remaining upcoming elections.

The current macro environment calls for action. How can the public and private sector collaborate effectively to secure democratic values and protect marginalized groups?

The Threat Of Deepfakes

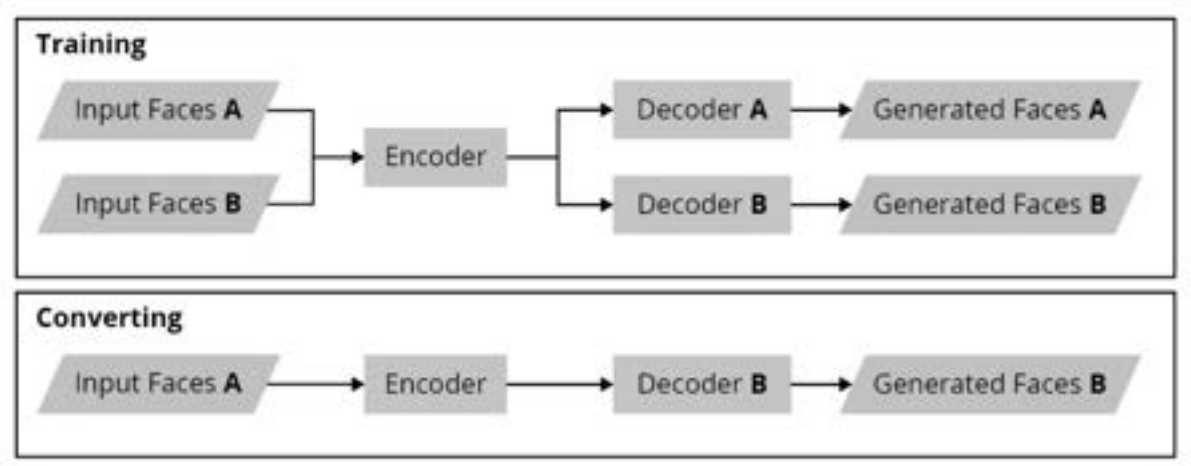
An Introduction To Deepfake Technology

Methods for manipulating media identities have been around for many years. It is widely known that images can be altered through various techniques such as photoshop. However, historically, producing high-quality manipulations of dynamic media was a time-consuming process. In contrast, creating deepfakes has never been easier or faster.

Advancements in AI have significantly simplified this task, enabling the creation of high-quality fakes with relatively little effort and expertise. These AI-driven methods, often utilizing deep neural networks, are commonly referred to as 'deepfakes'. These models are based on neural networks, which mimic an architecture inspired by the information processing of neurons in our brains.

Neural network models are built by stacking layers of mathematical functions called regression functions. Each layer takes in some input data and performs mathematical operations on it to transform it into a more useful form. These operations are used to analyze the relationship between input and output variables.

As shown in the figure below provided by the Carnegie Mellon University, creating a deepfake can be captured in a step-wise process.



Deep learning (DL) models such as autoencoders use layers to create an abstract representation of data, termed as latent representation. Encoding translates original data into this representation, while decoding reverses the process.

Additionally, Generative Adversarial Networks (GANs) represent another prominent tool for creating realistic deepfake images and videos. GANs consist of two neural networks, a generator and a discriminator, that are trained together. The generator creates synthetic data, while the discriminator evaluates its authenticity. Through this adversarial process, GANs improve over time, producing highly realistic media.

Moreover, Large Language Models (LLMs) are also used but to generate human-like dialogue as part of the storyline of the deepfaked media content. Other important technologies include: face swap algorithms that combine traditional computer vision with DL to replace faces in videos; Convolutional Neural Networks (CNNs) for detailed image recognition and manipulation; Recurrent Neural Networks and Long Short-Term Memory Networks to maintain temporal coherence in videos. By integrating these technologies, it is possible to create deepfake videos where the visual and auditory components are synchronized seamlessly. This results in highly convincing and coherent deepfake videos that can span long durations without losing consistency in narrative or visual quality, thereby broadening the potential applications and implications of deepfake technology across various public and private domains (Mitra et al. 2024).

An Overview Of Current State Of Deepfakes

A recent study contextualizes the current use cases of targeted deepfake manipulations. According to the cybersecurity company "Home Security Heroes", deepfakes have become a major concern in the digital realm. The report published in 2023 highlights its evolving capabilities and upcoming threats:

- The total number of deepfake videos online in 2023 is 95,820, representing a 550% increase over 2019.
- Deepfake pornography makes up 98% of all deepfake videos online.
- 99% of the individuals targeted in deepfake pornography are women.
- One in every three deepfake tools allow users to create deepfake pornography.
- It now takes less than 25 minutes and costs $0 to create a 60-second deepfake pornographic video of anyone using just one clear face image.
- Deepfake pornography is characterized by a select group of individuals who find themselves targeted for manipulative and often malicious purposes. These individuals, often public figures or celebrities, bear the brunt of deepfake creators' efforts.
- South Korean singers and actresses constitute 53% of the individuals featured in deepfake pornography and are the most targeted group.

This study demonstrates that deepfakes are also problematic in contexts other than politics and that the majority of deepfake use cases are targeted towards females. Overall, according to the UN around 57 percent of women worldwide have already been victims of image or video abuse on online platforms. This trend is likely to continue as the software required for generating deepfakes is freely available, open source, and easily accessible for download. A broad range of applications in the realm of deepfake pornography already exists. Apps like the Deepnude AI, Nudifier App and Undress AI allows its users to create images which leads to deceptively real non-consensual deepfakes.[1]

In the context of upcoming elections, such as in the EU and US, FemAI is posing the question of how targeted deepfake campaigns will affect female candidates. As seen in the study conducted by Home Security Heroes, South Korean singers and actresses account for over half of all measured deepfake targets. We pose the question if and how this picture will change during the super election year 2024 from the perspective of female politicians without ignoring the risk for public figures from the media industry.
Taking these data as key insights and analyzing the recent election in South Korea will deepen the research on the gendered nature of deepfakes and its impact on politics and elections.

AI And Elections: Lessons From South Korea

Recent developments in Gen AI have raised concerns as highly realistic images and videos flood the digital landscape. Although more than half of deepfake victims are from South Korea, the country managed to limit the impact of AI-generated deepfakes during the recent National Assembly elections. South Korea has faced numerous disinformation challenges, including domestic political rivalries, foreign interference from North Korea and China, and recent deepfake incidents targeting political figures.

Less than two weeks before South Korea's parliamentary elections on April 10th, 2024, authorities had pledged to take firm action against those attempting to sway the outcome using AI-driven deepfakes or cyberattacks. Shortly before the official start of campaigning on March 28, the National Police Agency (NPA) announced increased monitoring to prevent any unfair attempts to undermine the election's legitimacy. Even before the

[1] Because FemAI does not support the market for deepfake creation tools for non-consensual purposes, we have chosen not to provide a link.

campaign officially began, police and election officials were actively addressing criminal activities related to the vote.

To address the cyber threats to the electoral process, South Korea had revised the Public Elections Law to ban election-related deepfakes. Next to this, the government had initiated private sector efforts like content monitoring and AI-generated material watermarking in response to potential deepfake manipulations.

To protect the integrity of elections from evolving technological threats like deepfakes, it is essential to stay alert and use flexible strategies. There are two key takeaways for other countries that can be taken from this observation:

1) Public-private collaboration in managing risk in AI is key.
2) Combating disinformation with targeted and careful regulation to prevent misuse for political purposes needs to be adopted quickly.

However, the findings of deepfake for the group of singers and actresses as the primary targets of deepfake manipulation is often absent in most government actions and overlooked in discussions about solutions. With the upcoming elections in mind, the question arises to which extend politicians from marginalized groups will be potential victims. Therefore, in this white paper, FemAI will explore the effectiveness of deepfake detection tools through an intersectional feminist lens for female politicians.

A Perspective On Deepfakes Under The EU AI Act

Based on a policy paper provided by the EU in 2021, effective deepfake regulation can be targeted on five levels: Technology, Creation, Circulation, Target and Audience. In their policy paper "Tackling deepfakes in EU policy" , the EU suggests a list of seven ways to address the spread of deepfakes. When comparing the approach from 2021 with the current regulatory framework in 2024, the most relevant policy angles are (1) the recently passed EU AI Act (2), DSA and (3) the GBVD. This demonstrates the need for interlocking regulation frameworks to address the complexity of the threat of deepfakes for marginalized groups by applying a horizontal approach. The fast-paced technological development places policymakers under pressure to regulate with future developments in mind. Without ignoring the DSA and workstreams around the GBVD, this white paper focusses on how effectively the EU AI Act protects marginalized groups from the threat of deepfakes.

Under Article 52(3) of the AI Act, creators of deepfakes must disclose their artificial origin and provide information about the techniques used. This transparency aims to empower consumers with knowledge about the content they encounter and make them less susceptible to manipulation. However, transparency alone may not be sufficient to address the malicious potential of deepfakes, especially if creators find ways to circumvent disclosure requirements. Since the EU AI Act will not become binding until at least 2026, the urgency of regulating deepfakes won't be addressed by the upcoming EU regulation early enough.

Deepfakes are currently classified as "limited risk" AI systems under the AI Act, facing fewer regulations compared to "high-risk" systems. However, considering their significant harmful impacts, it can be argued they should be classified as high-risk. The AI Act also lacks a clear framework for legal liability for developers of deepfake technology, emphasizing preventative measures rather than punitive ones.

The establishment of the EU AI Office in February 2024 marks a significant step in promoting responsible AI practices within the European Union. One of the key functions of the AI Office is to encourage and facilitate the development of codes of practice at the Union level. In doing so, the aim is to effectively implement obligations related to the detection and labeling of artificially generated or manipulated content, tailoring regulations to protect marginalized groups. Under this mandate, the Commission is empowered to adopt and implement acts to approve these codes of practice, ensuring they meet certain standards and effectively address the challenges posed by artificially generated or manipulated content. However, there are still controversial issues regarding

clarity, specificity, and enforcement mechanisms within the AI Act. Whether the working groups compiling codes of practices will be able to fill the gaps of deepfake regulation is a question that still needs to be answered.

Advocates suggest criminalizing deepfakes for end-users to mitigate their harmful impacts. Also, enforcing legal consequences against individuals who create or disseminate them with malicious intent is required. The seven ways to regulate deepfakes posed by the EU in 2021 can be used as a guideline to foster a more effective deepfake regulation framework. Additionally, software developers and distributors should be mandated to incorporate measures to prevent the generation of harmful deepfakes through their products.

From a European perspective and with special regard to marginalized groups, the current deepfake regulatory approach is not protecting the spread of deepfakes effectively enough as policy suggestions such as watermarking ignore the use case of targeted, hostile deepfake campaigns. Public-private collaboration in regulating AI is key. The development of codes of practice must include technological perspectives and civil society representation to address gaps between them sufficiently. The challenge of the speed and virality of targeted deepfake campaigns, as seen in the Taylor Swift deepfake case in January 2024, remains mostly unsolved. Combating this type of harms with targeted and careful regulation can also help prevent the use deepfakes in political contexts and therefore needs to be adopted very quickly.

Why Are Deepfakes Problematic?

Deepfakes inflict harm on two profound levels: they violate individual rights and they undermine societal integrity. For the individual, the creation of non-consensual (especially sexual) content can cause deep emotional, psychological and reputational damage. On a broader societal level, deepfakes can be used for political manipulation, and the spread of false beliefs can lead to the erosion of trust in public institutions.

While concerns for deepfakes are pronounced, they are not uniform and vary in different contexts. It exhibits diverse implications across social media, public institutions and enterprise, each presenting a distinct set of consequences. The types of harm caused by deepfakes are also not novel, but the technology amplifies these issues by making them easier and faster to perpetrate. The acceleration of technological progress and the ease of causing harm risks normalizing illicit social manipulation, negatively impacting individuals and society at large.

In the case of an individual, whose face is seamlessly swapped with another's without their consent, it primarily threatens their privacy and leads to personal violations and abuses. As mentioned in the previous section, deepfakes pose significant individual-level harms, particularly for women, by enabling gender-based violence and eroding autonomy. They can also lead to the objectification of an individual, stripping them of their individual autonomy and reducing them to mere objects rather than whole people. Panoptic gaslighting, a severe form of emotional and psychological manipulation, can also manifest through deepfakes. This method of gaslighting leverages the widespread availability and sharing of videos on social media to create a pervasive and persistent form of manipulation. Over time, this can lead to a severe erosion of the victim's trust in their own memories and cognitive faculties, causing significant psychological distress and identity disintegration. Deepfakes, characterized by their falsified audiovisual nature, also present a significant threat in perpetuating financial frauds and identity theft of individuals.

Deepfakes raise concerns about their believable nature and the potential for harm, particularly in the context of political manipulation and media trust . This year, potentially marked by the most significant impact of AI on elections, deepfakes can play a major role in harming voters, candidates, and electoral integrity, necessitating multi-stakeholder interventions such as education, verification, and publicity moderation.

Deepfakes can cause "illocutionary harm" by forcing individuals to respond to fabricated statements or rumors, thereby undermining their communicative agency. This can occur even if the deepfake is debunked or not widely believed, as the necessity to publicly deny the falsehoods itself constitutes a harm. There have been historical instances where regimes compelled false confessions, illustrating the broader implications of this type of

wronging. When timed strategically, the spread of deepfakes – which can be difficult to detect and refute quickly – has the potential to distort public perception and trust. This erosion of trust has profound implications, potentially destabilizing democratic processes and undermining confidence in media and official communications. Furthermore, the proliferation of deepfake technology facilitates criminal activities, such as identity theft and fraud, by enabling the creation of fake identification documents. This not only threatens individual security but also poses broader economic and security risks

It is well established through prior research and through the content of this white paper that deepfakes disproportionately harm women and are being deployed as a new method for gender-based violence, eroding women's autonomy in online and offline spaces. Until now, deepfakes have primarily targeted public figures, especially women in the entertainment industry, a recent notable case being that of Taylor Swift. However, this issue is rapidly escalating, with deepfake victims now emerging in schools, workplaces, and domestic settings, exacerbating the existing challenges faced by women and other marginalized communities. This directly points to how this issue poses a significant threat to women's participation in democratic processes, particularly in elections. This has been a global issue, with politicians such as U.S. Congresswoman Alexandria Ocasio-Cortez and female and transgender leaders in countries like Bangladesh and Pakistan have been subjected to deepfakes. These malicious creations have been detrimental to their political careers and have even resulted in death threats in religious and conservative societies. In the European context, deepfake videos of the Le Pen family in France, Italian Prime Minister Giorgia Meloni and Dutch politicians and celebrities are a grim reminder of the far-reaching implications of this problem. This technology is often weaponized to create non-consensual explicit content, disproportionately targeting female politicians and individual voters to deter them from public life.

Such attacks not only inflict severe emotional and reputational damage but also deter women and other minorities from entering or continuing in politics. The pervasive fear of being subject to deepfake abuse, objectified and dehumanized repeatedly creates a significant barrier for females in politics who are already underrepresented in political discourse. This undermines gender equality in political representation, empowers misogynistic ideas and weakens democratic institutions.

Landscape Review Of Deepfake Detection Tools

The accelerating advancement of AI has rendered initial deepfake detection strategies less effective. The wider availability of deepfake databases and tools has made it easy for both experts and beginners to create realistic deepfakes. With the rapid spread of content on social media, these convincing deepfakes can be viewed by millions at scale and speed. As a result, it has become increasingly imperative to review and promote the best detection tools in this space in order to help protect individuals and societies globally.

The development of deepfake creation and detection algorithms has made significant progress, but challenges remain in terms of the threat posed by deepfakes to both individual privacy and national security. These issues are intensified by varying levels of awareness and concern among the public, as well as a lack of confidence in technology's ability to address the problem

In this section, the authors provide a comprehensive review of deepfake detection tools which can identify and mitigate the risks in different contexts and types of media including audio, video and audiovisual combined.

In this landscape review we aim to achieve a high-level analysis while maintaining conciseness and impact. To do this, we opted to categorize the detection tools into four distinct types based on their approaches. This methodology allows us to provide a broad yet insightful overview, highlighting key insights applicable to individual tools without delving into the technical intricacies of each one. By focusing on these categories, we can offer conclusions and recommendations that are relevant beyond the specific tools included in our review, providing a more holistic understanding of the current landscape in deepfake detection. The four categories are as follows:

1. **Visual, Temporal, and Forensic Analysis Tools**:
   These tools analyze visual and temporal data in media files to detect inconsistencies or anomalies and scrutinize the digital artifacts left behind during the creation or manipulation of media through forensic analysis. This includes examining frame-by-frame changes, looking for artifacts in image or video compression, and utilizing metadata analysis to ascertain the originality of the content.
   - **Mechanism**: These tools scrutinize visual and temporal elements of content, such as pixel inconsistencies, lighting mismatches, and timing irregularities. This also includes examining compression artifacts, noise patterns, and inconsistencies that may arise from frame manipulation. By analyzing the visual and temporal coherence of content, these tools can detect discrepancies indicative of deepfakes.
   - **Example**: NoisePrint provides a robust method for deepfake detection by extracting unique noise patterns from digital images and videos. This can help identify the source camera and the noise introduced during the digital image acquisition process.
   - **Detection tools and approaches:** NoisePrint, F3-Net, Capsule-forensics

2. **Physiological, Biometric, and Behavioral Analysis Tools**:
   This category focuses on detecting deepfakes by analyzing human physiological and behavioral cues that are difficult to replicate accurately with AI-generated content. This category also includes audio-visual synchronization issues between lips and spoken words and other mismatched audio-visual signals.

   - **Mechanism**: These tools analyze physiological signals (e.g., heart rate), biometric features (e.g., facial landmarks), and behavioral patterns (e.g., eye blinking) to detect abnormalities.
   - **Example**: Intel's FakeCatcher is a physiological analysis tool which detects deepfakes by analyzing subtle variations in blood flow patterns within video pixels to identify tampering.
   - **Detection tools and approaches:** DeepRhythm, EyeBlinking, DeeperForensics-10, Deepware Audio, LipForensics

3. **Provenance and Integrity Verification Tools**:

   These tools trace the origin and integrity of digital content. They use techniques such as identity watermarking, reverse image searching, and analysis of editing history to verify if content has been altered from its original state.

   - **Mechanism**: These tools focus on verifying the origin and integrity of digital content through metadata analysis and cryptographic techniques. By embedding and later verifying cryptographic markers, these tools ensure that the content has not been tampered with since its creation.
   - **Example**: Amber Authenticate integrates real-time monitoring, cryptographic watermarking, and machine learning to fingerprint videos and detect manipulations.
   - **Detection tools and approaches:** TruePic, Amber Authenticate, Serelay

4. **Hybrid Approaches**:
   Hybrid tools combine multiple detection methods from the categories listed above to improve accuracy and robustness. By integrating diverse analytical techniques, these tools aim to cover a broader range of deepfake generation methods and mitigate specific weaknesses inherent in single-method tools.

   - **Mechanism**: These tools combine multiple detection methods, including machine learning, digital watermarking, and forensic analysis, to improve accuracy and robustness. By leveraging various techniques, hybrid approaches can cross-verify content integrity and identify inconsistencies that single-method tools might miss.
   - **Example**: Sensity AI is a leading provider of hybrid deepfake detection solutions that combines visual, biometric, and audio analysis to detect deepfakes.
   - **Detection tools and approaches:** Deepware, Microsoft Video Authenticator, Reality Defender, Sensity AI, ResNet, DeepFake-o-meter

It is important to highlight the use of ML and DL (Machine Learning; Deep Learning) across all four categories of deepfake detection tools. Since deepfakes themselves are generated using AI techniques, such as GANs and other DL models, it is effective to use these technologies to combat them. Detection tools can identify the subtle inconsistencies and artifacts that other methods might miss. This approach also enables greater adaptation and resilience in the face of rapidly evolving deepfake generation methods. It is a potent example of fighting AI with AI.

Guiding Principles For Best Practice

The principles in this section serve as evaluation criteria to judge the four categories of deepfake detection tools. They are designed to address a wide variety of concerns ranging from technical robustness to social and environmental concerns to internal corporate practice. FemAI's intersectional feminist approach has informed the formulation of these principles and therefore center the needs of marginalized communities. Before articulating the principles, we will first describe the context in which the best tools must perform effectively.

Our perspective on the best deepfake detection tools is that they ought to work in both pornographic and political contexts, as such the best tools should identify deepfakes that are of both high-profile individuals (eg. politicians) as well as lesser-known people, of whom there is less (or no) training data available. They should detect deepfakes in audio-only, video-only and audio-visual media: political deepfakes can be harmful even if audio-only, while pornographic deepfakes are more visually oriented, and both will likely be audio-visual in most instances. The best tools will integrate with online platforms and social media where deepfakes can be distributed, and so they must work at scale and at speed. Lastly, the best tools will detect deepfakes found across the world, affecting a multitude of distinct cultures, communities and identities. Individuals must be protected from deepfakes no matter who they are or where they live.

Building on this, we have identified the four key principles of best practice for companies developing deepfake detection tools. In adhering to these principles, we argue that the best tools will be operating ethically and effectively across both political and pornographic contexts. They are:

1. **Inclusivity**
2. **Robustness**
3. **Efficiency**
4. **Responsibility**

Inclusivity

The principle of inclusivity is used to ensure that deepfake detection tools serve everyone, that they do not discriminate or fail people from certain places or backgrounds. Deepfakes should be detected in content regardless of who is depicted. This is a core tenet of our intersectional feminist approach: we must consider the diversity of identities and experiences to ensure equitable protection against digital manipulation. This approach aims to enhance the accuracy and fairness of deepfake detection and to address the power imbalances that exist in technology development and deployment. Ensuring inclusivity means that everyone, regardless of their background, has access to the same level of security and trust in digital content, thereby democratizing the power of these advanced technologies.

Since the tool needs to be scalable and international, it should be effective globally across geographies, genders, ethnicities, languages, skin color, and disabilities. In particular, it should be able to detect deepfakes involving subjects from the global majority. One paper in 2021 has already identified how the "the widely used FaceForensics++ dataset is overwhelmingly composed of Caucasian subjects", leading to a 10.7% error rate between groups of various races and genders.

This underscores the need for diverse training data sets so that hitherto underrepresented communities are not failed by these detection tools. Any datasets that the tool leverages or interacts with ought to be diverse, mitigating any biases and ensuring a fair level of protection across societies.

The developers of the best tools will ideally be themselves a diverse team of people from various backgrounds, helping to reduce and mitigate biases internally. Secondly, a diverse team will find it easier to naturally

empathize and identify with more subsets of the population. Such a team will also hopefully stimulate greater innovation compared to homogenous groups, overall aiding the tool's development and effectiveness.

The best tools will work well against deepfake content based on high-definition imagery or recordings as well as those involving more low-resolution and grainy content. This will account for the divergence of technological distribution across societies, so that both high- and low-quality content-based deepfakes can be detected.

Additionally, to help democratize access to the tool and ensure a greater distribution around the world, any user-interaction with the tool should be user-friendly to both laypeople and experts. Secondly, the tools should be available in as many markets as possible with fair, localized pricing to ensure just access.

Robustness

The tool needs to be technically sound and work as intended. This means that it is accurate, reliable and precise in identifying deepfakes with minimal false positives—in particular, it should not lead to more false positives for a specific group of people due to bias or poor design. Testing across diverse datasets ensures fairness and helps mitigate unintended biases. Additionally, the tool should not be based on racist or otherwise problematic theories or practice, such as phrenology, or affect recognition – shown to be inaccurate and unreliable across different cultures and identities.

The best tools will score highly on benchmarks such as F1 scores, which measure their precision and recall, and have a high 'Area Under the Curve' (AUC) value, reflecting their ability to effectively distinguish between real and fake content. Furthermore, robust tools will demonstrate high performance on standardized datasets like FaceForensics++ and in the Deepfake Detection Challenge (DFDC), which simulate a range of manipulative techniques.

While these benchmarks are valuable for evaluating the tool's accuracy, following the previous criticism of FaceForensics++, it is imperative to benchmark the tool against more diverse datasets to ensure scalability and adaptability.

The ability to update and learn from new data ensures the tool remains effective against the latest deepfake methods, making it suitable for real-world applications where rapid and accurate detection is essential.

To ensure longevity and futureproofing, the tool should employ methods that resist technical workarounds or hacks. The developers should ensure that any important datasets are kept up to date to keep the tool fully functional and accurate. The use of red teaming to simulate attacks or evade detection is encouraged in order to aid the tool's development by identifying weaknesses, leading to continual improvements and greater effectiveness.

Efficiency

The principle of efficiency ensures that deepfake detection tools operate effectively in high-demand online environments and at scale, e.g. on social media. An efficient tool must be capable of handling vast amounts of data swiftly, without the need for manual intervention, making platform integration necessary for the best tools. Automatic processing allows for real-time detection and response, essential for social media platforms and news outlets where the rapid spread of deepfake content can have immediate and widespread consequences.

Moreover, efficiency also encompasses the environmental and computational aspects of the tool's operation. It is crucial that the tool not only performs tasks quickly but does so using minimal computational resources. This

reduces the environmental impact associated with the energy-intensive processes of e.g. training and running AI models, aligning with growing concerns about AI sustainability.

An efficient deepfake detection tool, therefore, strives to balance between two needs: light resource consumption while maintaining high performance, ensuring it remains viable even as it scales across different platforms and user bases.

In practice, this means the tool should leverage advanced, yet resource-efficient, optimized algorithms that minimize the computational load without compromising detection capabilities. This approach ensures the tool can be deployed widely, including in regions with limited technological infrastructure, thereby broadening its accessibility and impact. The ongoing optimization of algorithms to reduce computational intensity is key to maintaining the best tools' efficiency over time.

Responsibility

The principle of responsibility is central to the ethical development and deployment of deepfake detection tools, embodying a commitment to continual improvement and realignment with the evolving needs of diverse communities affected by deepfakes. This responsibility entails identifying and addressing the technical challenges of deepfake detection while understanding and empathizing with the survivors of deepfakes from various backgrounds. By integrating feedback mechanisms, the tool developers can maintain an open dialogue with users and survivors, fostering an environment of continual learning and adaptation to new threats.

Accountability is a critical aspect of responsibility, which includes implementing mechanisms that ensure the tool and its developers are held to high ethical standards. This involves committing to transparency where methodologies and limitations of the tool are openly shared with the public to build trust and understanding. Additionally, the developers should employ comprehensive bias mitigation strategies that go beyond technical fixes to include ongoing internal training and education, ensuring that all team members are aware of and actively working to reduce biases in both the tool's operation and in their own work.

Explainability is vitally important too. It is essential for users—not just technical experts—to understand how the tool works and how it comes to certain decisions or judgments. This clarity empowers users and victims of deepfakes to critically evaluate and provide specific, constructive feedback on the tool, leading to more targeted and effective improvements.

By making the tool's output explainable, it allows a broader range of stakeholders to both understand it and contribute to its refinement and efficacy. This approach enhances the tool's capabilities and solidifies its role as a responsible actor in the fight against digital deception.

Application Of The Principles

In the following analysis, we systematically apply the four principles of best practice—Inclusivity, Robustness, Efficiency, and Responsibility—to the categories of deepfake detection tools. Our goal is to evaluate each category's alignment with these principles, providing a related score, to guide the identification of the most effective methodologies for combating deepfakes. The score is on a scale of **Low–Moderate–High**, where a **Low** or **Moderately Low** score indicates that the category of tool does not satisfy the principle in question, whereas a **High** or **Moderately High** score does.

Since some aspects of these principles pertain to organizational conduct and individual tool design—areas outside of this section's scope—our focus will primarily be on intrinsic attributes of each tool category. This allows us to concentrate on the technical and functional characteristics of the tools, setting aside considerations

related to specific corporate practices or the nuances of individual tool design or implementation. We will identify a best-in-class example to demonstrate such a detailed analysis, i.e. related to that specific organization and tool design, later in the paper.

(Readers will find a more thorough and detailed analysis, explaining the reasoning for the score under each principle, in the Appendix.)

Category name	Visual, Temporal, and Forensic Analysis
Examples	F3Net, NoisePrint, Capsule-forensics
Inclusivity	**High**
Robustness	**Moderately high**
Efficiency	**Moderately high**
Responsibility	**High**
Overall	Since this category assesses content characteristics without relying on demographic data, it has a broad applicability across diverse identities. These tools are robust, efficiently processing large datasets to detect a wide range of deepfakes, though they face challenges with newer, sophisticated evasion techniques and lower-quality media. With high responsibility, they provide clear, understandable outputs that explain detection decisions, enhancing user trust and minimizing biases.

Category name	Visual, Temporal, and Forensic Analysis
Examples	DeepRhythm, EyeBlinking, DeeperForensics-10, FakeCatcher, SyncNet, LipForensics
Inclusivity	**Moderate**
Robustness	**Moderate**
Efficiency	**Moderate**
Responsibility	**Moderate**
Overall	These tools aim to minimize demographic biases by analyzing universal human traits, but their effectiveness depends heavily on training with diverse datasets. They are generally robust against deepfakes that poorly mimic human behaviors but falter with non-human elements and low-quality inputs. They may also struggle as deepfake technology evolves and become better at simulating human behavior and traits. Their broad potential use is hindered by high computational demands, significant privacy concerns, and challenges in transparency and understandability.

Category name	Provenance and Integrity Verification
Examples	Truepic, Amber Authenticate, FotoForensics, Serelay
Inclusivity	**Moderately low**
Robustness	**Moderately low**
Efficiency	**High**
Responsibility	**High**
Overall	These tools excel in handling diverse content impartially since they do not rely on demographic data, making them less prone to cultural or physical biases. They are known for their robustness in detecting signs of media manipulation through artefacts and metadata, though they may struggle against highly sophisticated deepfakes that do not exhibit signs of watermarking. These tools are robust in identifying a limited range of media manipulations but are efficient and transparent, requiring low resources while providing clear evidence of alterations. Most importantly though, they are not applicable to most deepfake content unless they have been watermarked or authenticated and so are limited in their scope and scale.

Category name	Hybrid Approaches
Examples	Deepware Scanner, Microsoft Video Authenticator, Reality Defender, Sensity AI, DeepFake-o-meter, ResNet-50
Inclusivity	**High**
Robustness	**High**
Efficiency	**Moderately high**
Responsibility	**High**
Overall	Hybrid approaches in deepfake detection excel across all key principles by leveraging a combination of detection methods to enhance overall performance, scalability and adaptability. These tools effectively address the diverse challenges posed by deepfakes, providing robust, scalable, and responsible solutions that perform well in a variety of settings. While they can be highly inclusive and robust when trained on diverse datasets, their inefficiency during training phases is problematic and the challenge of transparency due to AI's opaque nature requires careful management and ongoing refinement.

Best In Class

To highlight a particularly effective deepfake detection tool, *Reality Defender* performs well across the four principles of best practice and as an example of the 'Hybrid Approach'. It advertises it's "patented multi-modal approach" that leverages a variety of neural network models such as Convolutional Neural Networks (CNN), Transformers, and Vision Transformers (ViTs) and performs "spatial, temporal, and frequency domain analysis along with domain specific feature losses (such as artifacts in images)" as well as uses "generative content fingerprinting technology". Its hybrid approach enhances its effectiveness across diverse types of content and demographics. The key for the company is to remain committed to continuous improvements and updates, using diverse datasets, and ensuring their internal culture adheres to the principle of Responsibility, centering the needs of marginalized peoples. We will now unpack the analysis of the tool according to the four principles.

Inclusivity

Reality Defender exemplifies a high degree of inclusivity through its use of diverse datasets and a multi-modal approach that accommodates various types of content such as videos, images, audio, and text. The CEO, Ben Coleman, claimed that they incorporate "a wide variety of accents, skin colors and other varied data" into its training datasets. This approach helps ensure that the tool is effective across a wider range of demographic groups and media types, reducing the potential for bias that is often present in tools trained on more homogeneous datasets. The inclusion of telephone-quality audio analysis further indicates its capability to handle varied media qualities, which is particularly important for ensuring effectiveness in less technologically advanced regions.

As far as we can see there is no public evidence of the tool's accuracy across a wide range of demographics (ethnicities, skin color, age, etc.), but the statement by the CEO at least acts as a statement of intent towards protecting diverse communities.

One solution would be to enable third party auditors to assess their training data and certify the effectiveness in detecting deepfakes depicting a diverse variety of people from around the world and across societies.

Robustness

The robustness of *Reality Defender* is underscored by its integration of multiple neural networks, (CNNs, Transformers, and Vision Transformers (ViTs)). This range of models helps enhance the tool's ability to detect subtle nuances of digital manipulation. Including spatial, temporal, and frequency domain analyses lead to a more comprehensive scrutiny of media files, allowing the tool to maintain high accuracy and reliability even as deepfake technologies evolve. However, the robustness of this approach is only as good as the commitment to ongoing refinement and improvements to keep up with the pace of change in the field of deepfake technology.

There is however scant evidence of any benchmarking to back up it's claims of being best-in-class, though the CEO posts that they "target 95% accuracy with all public and proprietary training sets." Ideally the company would publish their test results or, similarly, allow a third-party auditor to certify their claims.

Efficiency

Reality Defender appears particularly efficient in its offering of platform-agnostic APIs, designed to operate effectively integrated with online platforms, and a web app. The tool's ability to process large volumes of data in real-time is particularly important for social media monitoring and broadcast media verification. However, the computational demands are significant due to the complexity of the analyses involved and the use of neural networks. The CEO does claim that they are "optimizing compute requirements by isolating only the part of the media that requires scanning" which hopefully reduces the environmental impact. On this point, the overall

approach suffers negative criticism compared to other less computationally intensive methods, given that they use multiple AI models.

Responsibility

Reality Defender can be touted for its transparency and explainability. The tool provides clear probabilistic assessments of media authenticity, which are easily understandable, fostering trust and reliability among users. Furthermore, the absence of reliance on watermarking, opting instead for an inference system that does not require any ground truth, as watermarking does. They also demonstrate that their approach respects user privacy and data protection by refraining from training the AI models with uploaded data and enabling users to delete data immediately after use.

Reality Defender provides a robust and adaptable solution in the landscape of deepfake detection. The continued effectiveness of *Reality Defender* hinges on continuous updates and the commitment to refining its techniques in response to evolving deepfake technologies. It must also build and maintain trust in its corporate culture and practice—a continued commitment to transparency and user privacy in its operations will aid in this.

Employing a variety of neural network models, *Reality Defender* conducts comprehensive analyses across multiple modalities. This approach supposedly extends its applicability across a wide array of demographics, supporting its efforts to address biases. However, without public evidence confirming this, the aforementioned solution of third-party audits to certify its effectiveness is called for by this paper. This would strengthen its fulfilment of the principle of Inclusivity and maintain its Robustness in the ever-evolving landscape of deepfakes.

Conclusion

The current spread of deepfakes is a global concern. According to the UN, 57 percent of women worldwide have already fallen victim to image or video abuse on online platforms. Deepfake videos in 2023 increased by 550% when compared to 2019 with deepfake pornography constituting 98% of all deepfake videos online. 99% of the individuals targeted in deepfake pornography are women. Neither regulation nor the deepfake detection tool landscape is committed enough to solve the gendered nature of these harms effectively.

The harms created by deepfakes are real and pose significant challenges at both individual and societal levels. At an individual level, deepfakes infringe individual privacy and enable gender-based violence. On the other hand, at a broader societal level, deepfakes facilitate political manipulation and trust erosion in public institutions, exacerbating issues of disinformation and undermining democratic processes. When viewed from a gendered perspective, especially in the super election year 2024, deepfakes disproportionately harm female politicians and candidates from other marginalized communities. This can be a huge deterrent for achieving gender equity in politics and discourage females and other marginalized communities to participate in public life.

In pursuit of finding the best-in-class deepfake detection tool available, we analyzed the existing deepfake detection tool landscape. This led us to identify four categories of tools, each employing different methods to detect deepfakes, and a set of guiding principles for best practice that we used to evaluate these tools. These four principles – inclusivity, robustness, efficiency, and responsibility – if followed, help ensure effective and ethical deployment of deepfake detection tools. It is our opinion that platforms (including those pornographic in nature), media outlets, and social media should integrate such tools into their content moderation practice and protect individuals and societies around the world. By judging tools on these four principles, they will make informed decisions that will not only benefit the most marginalized in society but will lead to the best and most accurate results.

The white paper acknowledges the fact that we are all vulnerable to deepfakes. But the current statistics demonstrate that they impact unequally across gender lines and that many detection tools can fail to protect the already marginalized. In conclusion, addressing the complex challenges posed by the spread of deepfakes targeting females demands urgent and comprehensive action. There are, however, better tools that are more ethical and effective, so platforms and regulators can protect their users and citizens. The regulators must compel platforms to detect and take down deepfakes, not just at election times, but every day and in other contexts that impacts ordinary lives. We have proposed four principles to help them choose adequate providers and to help tool developers protect us all.

Additionally, mandating prevention measures such as self-commitment strategies for software developers and distributors as well as establishing effective enforcement mechanisms, can serve as methods to combat the global threat of deepfakes. It is also important that there is a global collaboration among public and private sector actors to protect democracies. This process, such as the development of a code of practice under the EU AI Act, must actively involve marginalized groups and civil society and should be designed transparently.

Moreover, initiatives to promote AI literacy and collaborative awareness campaigns are essential to empower the public. By addressing these challenges comprehensively, we can foster a safer digital environment and protect our society against the threats of deepfakes.

Applying an intersectional feminist approach has also helped us identify the limitations of this research. Most findings are considered in the binary gender system overlooking intersectional identities like ethnicity, age or social status. For deepfake detection tools to be effective for all, more research is required to study the impact of deepfakes on other marginalized communities and the compounding effects involved. We also advocate for

diverse training data sets which are currently absent in the design of these detection tools. Additionally, within the scope of this white paper, we do not address the added value and benefits of AI-generated synthetic media in the fields of education, healthcare and arts.

This white paper serves as a primer and suggests concrete next steps that range from the need for user-facing technical solutions, increased cross-sector collaboration (Policy, Research, Civil Society, Business) to a low-barrier step-guide on how to respond if targeted by a deepfake. As part of this white paper launch, FemAI invited the public as well as female politicians to include their perspective on this topic. We encourage to foster alliances on the threat of deepfakes within and outside the FemAI ecosystem. FemAI will continue efforts to provide the room as a think tank to execute further research integrating diverse perspectives while keeping an intersectional feminist approach. We emphasize the need for increased funding in this research field from the private and public sector. This is also a reminder for tech companies of their unwavering social responsibility even in an ever-evolving technology landscape.

Appendix

Category name	Hybrid Approaches
Examples	Deepware Scanner, Microsoft Video Authenticator, Reality Defender, Sensity AI, DeepFake-o-meter, ResNet-50
Inclusivity	**High**— • By combining multiple methods, these tools are theoretically effective across a wider range of content and depicted demographics. They can compensate for the limitations of one detection method with the benefits of another. • Assuming the use of diverse datasets for training, tools in this category can be very effective across various content types and demographics. They could therefore scale relatively effectively, if any biases are mitigated against. • As a result, the overall inclusivity of these tools depend heavily on the datasets used and model testing carried out.
Robustness	**High**— • Leveraging multiple techniques similarly strengthens the tool and leads to reduced false positives and increased resistance to attacks. • These tools often score very highly on standard benchmarks and are harder to evade since they don't rely on one method. Advanced models in this category can have strong performance, minimal false positives, and can be resistant to attacks. • They can also be adaptable to new types of deepfakes by being trained on new datasets. • There is a susceptibility to the use of adversarial attacks that mislead the tool through slight changes to the input data. Continual improvement and training is required to overcome this.
Responsibility	**High**— • Tools in this category can theoretically address bias more effectively by combining different detection methods that can cross-validate each other, reducing the chance of biased outcomes. Though bias mitigation strategies must always be implemented regardless. • Some also have some explainability and transparency features such as providing confidence scores in detections and signaling what aspect of media is flagged as likely manipulated. • On the negative side, AI models can still be opaque and therefore it can be difficult to otherwise explain and for users to understand their outputs, reducing their overall transparency. This can be overcome but must be prioritized and designed for.
Overall	Hybrid approaches in deepfake detection excel across all key principles by leveraging a combination of detection methods to enhance overall performance and adaptability. These tools effectively address the diverse challenges posed by deepfakes, providing robust, scalable, and responsible solutions that perform well in a variety of settings. While they can be

	highly inclusive and robust when trained on diverse datasets, their inefficiency during training phases is problematic and the challenge of transparency due to AI's opaque nature requires careful management and ongoing refinement.

Category name	Physiological, Biometric, and Behavioral Analysis
Examples	DeepRhythm, EyeBlinking, DeeperForensics-10, FakeCatcher, SyncNet, LipForensics
Inclusivity	**Moderate**— • Analysing physiological and behavioural signals have the potential to be less biassed towards certain demographics since they are theoretically universal. • However, this requires a diverse dataset to ensure it is trained to compensate for demographic variability in expression. • Various disabilities and health conditions (eg. strokes) that are likely not contained in training datasets, and will likely lead to inaccurate results, need to also be taken into account.
Robustness	**Moderate**— • Tools in this category use techniques that can be hard for deepfakes to circumvent, since they use specific metrics (eg. eye blinking patterns) that current deepfakes struggle to imitate well. • This approach is protected against novel deepfake techniques, so long as they continue to struggle to replicate human characteristics convincingly. However, these tools may struggle with deepfakes that do not alter physiological or biometric data (e.g. non-humans and background content). • They are also highly sensitive to the quality of input, such as poor video resolution or lighting and can suffer from higher error rates in noisy environments or when the audio quality is poor, for example.
Efficiency	**Moderate**— • These tools are potentially applicable without needing additional customization or manual intervention since they can be used across demographics and regions, but only if adequately trained. • However, they can be generally less efficient compared to other methods due to their high computational demands for real-time processing of complex, multi-dimensional human data, which does not reduce once trained. • Since these tools are also AI powered, they also run the added criticism of being resource-intensive and inefficient to train.
Responsibility	**Moderate**—

	• Tools can mitigate against biases through diverse training datasets, but they also pose significant challenges in terms of privacy and complexity: • They analyse sensitive human data, necessitating stringent privacy safeguards. Ensuring that data is handled and stored securely is crucial. The risk is in the processing and storing of biometric data (like facial recognition) and voice patterns. • Secondly, these tools often use algorithms that may not be transparent or understandable to non-experts. This could pose challenges in explainability, making it difficult for users to understand how decisions are made.
Overall	These tools aim to minimise demographic biases by analysing universal human traits, but their effectiveness depends heavily on training with diverse datasets. They are generally robust against deepfakes that poorly mimic human behaviours but falter with non-human elements and low-quality inputs. They may also struggle as deepfake technology evolves and become better at simulating human behavior and traits. Their broad potential use is hindered by high computational demands, significant privacy concerns, and challenges in transparency and understandability.

Category name	**Provenance and Integrity Verification**
Examples	Truepic, Amber Authenticate, FotoForensics, Serelay
Inclusivity	**Moderately low**— • These tools do not rely on demographic data, making them applicable across different groups without inherent biases related to physical or cultural characteristics. • However, it may not be comprehensive for all types of deepfakes, particularly if certain types of media from less developed regions are not adequately represented in training data or testing scenarios. • The tools will also not work with content that is not watermarked or authenticated in some way, therefore drastically reducing it's scalability.
Robustness	**Moderately low**— • This is a generally well proven method for verifying media authenticity with minimal false positives, but only in so far as there is a 'ground truth' or digital trail for them to follow. • Analysing the consistency of digital artefacts, metadata, or editing footprints, is generally reliable but can be bypassed by sophisticated techniques designed to remove or alter these traces. • This also means these tools are less effective if the content was never watermarked or authenticated in the first place because they rely on these markers to verify integrity.
Efficiency	**High**—

	• Such tools can be integrated into platforms with minimal intervention, since they quickly analyse metadata or use relatively straightforward algorithms to check for inconsistencies, making them suitable for quick assessments. • They also generally have lower computational requirements compared to more complex or machine learning approaches.
Responsibility	**High**— • Tools in this category can provide clear, understandable evidence about whether content has been altered. For example, it can highlight altered timestamps or inconsistent file structures. • This clarity can make it easier for non-experts to understand and trust the results. • Since these tools often analyse metadata and potentially sensitive information, implementing strong data protection measures is crucial. • These tools are also less dependent on demographic data, meaning they are less likely to inadvertently perpetuate or create biases.
Overall	These tools excel in handling diverse content impartially since they do not rely on demographic data, making them less prone to cultural or physical biases. They are known for their robustness in detecting signs of media manipulation through artefacts and metadata, though they may struggle against highly sophisticated deepfakes that do not exhibit signs of watermarking. These tools are robust in identifying a limited range of media manipulations but are efficient and transparent, requiring low resources while providing clear evidence of alterations. Most importantly though, they are not applicable to most deepfake content unless they have been watermarked or authenticated and so are limited in their scope and scale.

Category name	**Visual, Temporal, and Forensic Analysis**
Examples	F3Net, NoisePrint, Capsule-forensics
Inclusivity	**High**— • Since tools here focus on content characteristics and not on demographic features such as race, gender, or age, this ensures applicability across diverse subject identities. • There may be limitations in effectiveness when dealing with lower-quality media, which could disproportionately affect less technologically advanced regions.
Robustness	**Moderately high**— • These tools are capable of detecting a wide range of deepfakes due to inconsistencies and anomalies that are typical of manipulations. • However, their effectiveness can be challenged by new techniques that are designed to evade such detection mechanisms (or aren't included in the training dataset).

Efficiency	**Moderately high**— • This category is generally efficient, designed for quick processing of large datasets essential for real-time and high-volume environments. • Though, more complex forensic analyses may require more processing power and time, potentially reducing overall efficiency.
Responsibility	**High**— • The tools typically provide clear outputs that indicate why a particular piece of content was flagged as manipulated, like pixel inconsistencies or frame-by-frame changes, making them accessible and trustworthy to users. • These tools often do not directly involve demographic data, ensuring that the visual and temporal analysis algorithms are free from any form of contextual bias.
Overall	Since this category assesses content characteristics without relying on demographic data, it has a broad applicability across diverse identities. These tools are robust, efficiently processing large datasets to detect a wide range of deepfakes, though they face challenges with newer, sophisticated evasion techniques and lower-quality media. With high responsibility, they provide clear, understandable outputs that explain detection decisions, enhancing user trust and minimizing biases.


FemAI
CENTER FOR
FEMINIST
ARTIFICIAL
INTELLIGENCE