ORIGINAL ARTICLE

INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026

Journal: https://www.academicpublishers.org/journals/index.php/ijai

COMMERCIAL CHATBOTS DECEIVING DEEPFAKE DETECTORS: THE NAÏVE EXPOSURE OF GENERATIVE AI CAPABILITIES

Djurayeva Buvsara Abdumannonovna
Jizzakh state pedagogical university
Department of information technologies and systems
(PhD), Associate Professor

Abstract

This article is based on research conducted by Kim et al. (2026) and analyzes how the powerful capabilities of generative AI systems, presented through user-friendly interfaces, can fundamentally undermine state-of-the-art deepfake detectors. Rather than proposing a new manipulation technique, the study demonstrates how an ordinary user, using only standard prompts that do not violate safety guidelines and commercial generative AI systems, can circumvent the most advanced deepfake detection methods. Specifically, the researchers found that generative AI systems explicitly articulate authenticity criteria, externalizing them through unconstrained reasoning processes, and transform these criteria into reusable refinement objectives. As a result, the refined images simultaneously evade detectors, preserve identity verification by commercial facial recognition APIs, and maintain significantly higher perceptual quality. Most importantly, widely accessible commercial chatbot services pose a much greater security risk than open-source models, as their high realism, semantic controllability, and low-barrier interfaces enable even inexperienced users to achieve effective evasion [1].

Keywords

Large Language Models (LLMs), Text-to-image generation, Image manipulation, Detection evasion, API security, Face recognition systems

INTRODUCTION

In recent years, the rapid development of generative artificial intelligence (GenAI) technologies has led to a dramatic increase in the quality of deepfake content. Ordinary users now have the capability to create highly realistic fake images, audio, and videos without requiring complex technical knowledge. This situation not only raises concerns about personal privacy and reliability but also poses serious threats to democratic processes, national security, and the global economy[10,5].

One of the primary tools in combating deepfake technology is specialized detectors. These detectors are designed to identify content created using artificial intelligence. However, researchers note that a continuous "arms race" persists between increasingly sophisticated generative models and the detectors developed to counter them[2,4].

The study analyzed in this article reveals a new and unexpected twist in this race: instead of developing new generations of models, deepfake creators have found a way to bypass state-of-the-art detectors by utilizing the simple, legitimate capabilities of existing commercial chatbots. This new threat is termed "naïve exposure" and demonstrates the need to reconsider existing strategies in the fight against deepfake risks.

MAIN PART

1. The new threat concept: "Naïve exposure"

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03, 2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

Feature	Commercial Chatbots	Open-Source Models
Realism Level	Very high, continuously updated	Variable, often lower
Ease of Use	High, requires no specialized technical knowledge	Moderate, requires installation and configuration
Semantic Control	Strong, understands complex prompts	Limited
Risk Level	**High**	**Medium**

The Essence of the Problem. The core idea of the research conducted by Kim et al. (2026) is as follows: modern commercial chatbots (e.g., ChatGPT, Gemini, and others) provide users with powerful capabilities for analyzing and refining images. Utilizing these capabilities, the researchers developed a simple method for creating images that can bypass deepfake detectors [1].

The process consists of the following stages:

- **Initial deepfake creation:** Initially, a fake image is created using simple deepfake generation tools.
- **Detection by detector:** This initial deepfake is identified by modern detectors.
- **Refinement via chatbot:** The detected deepfake image is uploaded to a commercial chatbot, and simple prompts such as "make this image more realistic" or "remove the artificial signs from this image" are sent.
- **Evasion of detector:** The image processed and refined by the chatbot is no longer detected by deepfake detectors.

2. Why does this work?

The researchers found that generative AI systems possess internal concepts known as "authenticity criteria." These criteria encompass the AI model's knowledge of what makes an image real versus fake. The problem is that these criteria are externalized through the model's unconstrained reasoning processes, allowing users to transform them into reusable refinement objectives.

In other words, an AI model that understands the specific features deepfake detectors use to identify fakeness can "cleanse" the image by removing precisely those features. This represents a new, complex manifestation of the classic struggle between "sword and shield"[1].

3. Commercial Chatbots vs. Open-Source Models and Their Risk Level

One of the most significant conclusions of the research is that commercial chatbots pose a substantially greater security risk compared to open-source models. There are several reasons for this:

The high realism of commercial chatbots, their ability to understand complex language prompts, and their simple interfaces make them an ideal tool for inexperienced users. Specialized technical knowledge or complex software tools are no longer needed to create deepfakes and conceal them from detectors[1].

4. Identifying the Weakness of Deepfake Detectors Based on Scientific Research

The vulnerability of existing deepfake detectors has also been confirmed in several scientific studies. The 2026 International AI Safety Report notes that currently, no AI model can reliably detect high-quality deepfake videos [3,8]. The report assesses the effectiveness of deepfake detectors as "limited" and advises against drawing general conclusions about their success rate[5].

Furthermore, research on "Industrialized Deception" highlights that the competition between generation and detection in the struggle between generative AI and deepfakes

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

continues. As detectors improve, the methods used to deceive them also become more sophisticated. This indicates that the "arms race" is entering a new phase[2].

5. Social and Political Consequences of the New Threat

5.1. Threat to Democratic Processes. The proliferation of deepfake technology through such easy and effective tools enables serious interference in election processes. Spreading false information about candidates using fake videos and audios, influencing voters' opinions, and creating an atmosphere of distrust in society are all possible. As Ferrara (2026) notes, generative AI enables the creation of "synthetic realities" and can erode society's fundamental epistemic foundations (knowledge and belief systems) [5,10].

5.2. Personal Security and Fraud. Instances of identity theft and fraud through deepfakes could increase dramatically[8, 9] Crimes such as stealing money from bank accounts via fake voice messages, deceiving company executives, or inciting ordinary citizens to inappropriate actions could become widespread. Specifically, in India in 2024, losses amounting to 22,000 crore rupees (approximately 2.6 billion US dollars) were caused by cyber-frauds, most of which were perpetrated using AI[8].

5.3. Risks of "Epistemic Fragmentation" and "Synthetic Consensus". Research conducted by Loth et al. (2026) suggests that large-scale text generation can lead to systemic risks termed "epistemic fragmentation" and "synthetic consensus." This could result in the loss of a shared concept of truth in society, leading different groups to live within their own "alternative facts.[2]"

5.4. Crisis of Trust: The "Generative AI Paradox" Concept. According to the "Generative AI Paradox" concept introduced by Ferrara (2026), synthetic media becomes so widespread that societies rationally begin to doubt all digital evidence. This could lead to a crisis of trust in all areas, from the judicial system to everyday communication. If any video, audio, or document could be fake, proving genuine evidence becomes impossible[10].

6. Existing and Potential Solutions

6.1. Provenance and Watermarking Technologies. One of the most promising directions in combating deepfakes involves tracking the origin of content (provenance) and watermarking technologies. The C2PA (Coalition for Content Provenance and Authenticity) standard enables the cryptographic verification of a piece of content's creation and modification history[2]. Google's SynthId technology allows for the identification of images generated by generative models by adding a unique identifier during their creation process[7].

6.2. Multi-Layered Approach: The "Defence in Depth" Strategy. A single tool is insufficient in the fight against deepfake threats. Experts propose the "Swiss Cheese" model or the "Defence in Depth" strategy[7]. This approach involves combining several layers of protection:

- Technical detectors
- Provenance verification
- Platform governance and moderation
- User awareness and media literacy
- Redesigning institutional processes[10].

6.3. Human Factor and Education. While technical solutions are crucial, the human factor is also of decisive importance. Raising user awareness about deepfake risks, teaching them methods to verify content, and developing critical thinking skills are essential tasks. Simultaneously, the habit of verifying any suspicious content through multiple sources and consulting reliable sources must be cultivated[8].

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

6.4. Legal and Policy Measures. Legal measures against the deepfake threat are being taken at the international level. The European Union's Council of Europe Framework Convention on Artificial Intelligence (2026) requires AI systems to adhere to strict ethical standards, ensuring transparency and effective oversight throughout their lifecycle. The United States, the United Kingdom, Canada, and other countries are expected to sign this convention[10].

CONCLUSION AND RECOMMENDATIONS

The research conducted by Kim et al. (2026) has revealed a new and serious challenge in the fight against the deepfake threat: it is possible to bypass even the most advanced deepfake detectors using the simple capabilities of commercial chatbots. This situation necessitates a reconsideration of existing security strategies and the development of new, comprehensive solutions [1].

The following recommendations can be put forward:

1. **Continuous updating of deepfake detectors:** Detectors should be developed not only considering new generative models but also the "cleansing" capabilities of existing chatbots.
2. **Widespread implementation of provenance technologies:** Implementing systems to track content origin and verify it cryptographically should be made mandatory across all platforms.
3. **Strengthening international cooperation:** As the deepfake threat is a global problem, international cooperation, information exchange, and the development of unified standards are crucial in combating it.
4. **Developing user guidelines:** Creating practical guides for ordinary users on how to identify deepfake content and protect themselves from it.
5. **Security protocols for commercial chatbots:** Companies developing commercial chatbots must implement additional security measures to prevent their systems from being used for such purposes.

The fight against deepfake technology is a long and complex process that requires a combination of technical, legal, educational, and institutional measures. New threats like "naïve exposure" demonstrate just how dynamic and complex this struggle truly is.

REFERENCES

1. Kim, S., et al. (2026). Naïve Exposure of Generative AI Capabilities Undermines Deepfake Detection. *arXiv preprint arXiv:2603.10504*. Available at: https://arxiv.org/abs/2603.10504 [Crossref] [1]

2. Loth, A., Kappes, M., & Pahl, M. O. (2026). Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)*. Dubai, United Arab Emirates: ACM. Available at: https://arxiv.org/abs/2601.21963 [Crossref] [2] [8]

3. Ferrara, E. (2026). The Generative AI Paradox: GenAI and the Erosion of Trust, the Corrosion of Information Verification, and the Demise of Truth. *arXiv preprint arXiv:2601.00306*. Available at: https://arxiv.org/abs/2601.00306 [Crossref] [3] [9]

4. Talib, H. (2026, February 5). Deepfake Detection in the 2026 AI Safety Report. *Deepfake Field Notes*. Available at: LinkedIn publication [4]

5. Wig, T., et al. (2026, February 17). Experts Say No AI Model Can Reliably Detect Deepfakes Yet. *MediaNama*. Available at: https://www.medianama.com/2026/02/223-ai-enabled-cybercrime-india-deepfake-threat/ [5]

ISSN: 2692-515-x, Impact Factor: 9.23
American Academic publishers, volume 6, issue 03,2026
Journal: https://www.academicpublishers.org/journals/index.php/ijai

6. Reality Defender. (2026). 2026 Deepfake Outlook: Cautious Optimism as Deepfake Threats Escalate, While Detection Closes the Gap. *Reality Defender Insights*. Available at: https://www.realitydefender.com/insights/2026-look-forward [6]
7. Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. *arXiv preprint arXiv:2503.02857*. Available at: http://arxiv.org/abs/2503.02857 [4]
8. Bengio, Y., et al. (2023). Generative AI models should include detection mechanisms as a condition for public release. (Reference in Talib, 2026) [4]
9. Singh, T. (2026). AI-enabled cybercrime in India: Scale and impact. In *India AI Impact Summit*. (Cited in Wig et al., 2026) [5]
10. Maheshwari, R. (2026). Digital Personal Data Protection Act and AI governance. In *India AI Impact Summit*. (Cited in Wig et al., 2026) [5]