Deepfake Detection: Human Performance Versus AI Tools, A Comparison of Accuracy and Effectiveness

Anastasiia Iufereva[1] and Peter Mozelius[2]
[1]IULM University, Milan, Italy
[2]Mid Sweden University, Sundsvall, Sweden

anastasiia.iufereva@studenti.iulm.it
peter.mozelius@miun.se

Abstract: Image manipulation is a phenomenon much older than digital image handling and generative artificial intelligence (GenAI). In the digital era, researchers have made a distinction between cheapfakes and deepfakes. The creation of cheapfakes requires a relatively low technical editing level and does not depend on any GenAI technology. This study had a focus on deepfakes only and explored the role of GenAI in the generation of deepfakes as well as the role of AI in tools for detecting deepfakes. The term deepfakes refers to high-quality and synthetic media content created with the use of deep learning and generative artificial intelligence. Recent advances in deepfake generation have made such content even more realistic, making it harder to identify. The rise of AI-convincing content is a growing social issue that poses serious challenges for its detection. Although the importance of deepfake detection is widely recognized, research comparing the performance of humans and AI on deepfakes analysis is still in its early stages. This study addresses this gap by conducting a comparison between human analysis and AI-based detection tools to evaluate their accuracy and effectiveness in identifying deepfakes. The testing was based on a set of AI-generated images related to the Israel–Hamas war, which were circulated on social media during 2023-2024. Future research should focus on testing a broader range of AI-based detectors to evaluate their effectiveness against different types of deepfakes, such as video, music, audio, text, including less sophisticated AI-generated manipulations as cheapfakes.

Keywords: Deepfakes, AI, GenAI, AI-Detection, Detection tools, Image manipulation

1. Introduction

Image manipulation, defined as the alteration of photographs to change their appearance or content, is a phenomenon that stretches back to the earliest days of photography. In 1826 or 1827, the exact date is not verified, the French inventor Nicéphore Niépce preserved and presented the first heliograph photocopy captured by a camera obscura. A new technique for generating and storing images was created, with a quality that continuously improved in the 19th century (Mcquire et al., 2024). A new technique that also opened up new ways for image manipulation, not only for retouching scars and wrinkles on portraits, but also for political propaganda. A well-known example from the 19th century is when multiple negatives were combined to create fake images during the American Civil War (Palmquist & Kailbourn, 2005). As discussed in Pisch (2014), retouching techniques were frequently used in the Soviet Union during the Stalin era, for various purposes such as “smoothing Stalin's pockmarked face and removing litter from his path; inserting text on banners so that the idea becomes legible; enlarging an adulatory crowd through montage…" (Dickerman, 2000, p. 141).

Around 150 years later, when the first digital camera prototype was presented in late 1975 or early 1976, the exact date is not verified here either; a new era with computer-assisted manipulation was born. Digital image manipulation without involving any generative AI (GenAI) techniques is today known as cheap fakes or cheapfakes. The disinformation or misinformation in cheapfakes does not necessarily have to be cheap, and that cheap refers to the affordable tools and techniques that are used. Deepfakes, on the other hand, are created with the use of GenAI, and the term is built by combining the first part of deep learning with fake. In this context, deep learning refers to the meaning of deep learning for multi-layered neural networks, and not to the meaning of deep learning as used in pedagogy. Moreover, deepfakes today are multimodal, comprising modalities such as image deepfakes, audio deepfakes, and video deepfakes (Ahmed & Chua, 2023). This study had a focus on deepfakes only, with deepfakes definiens as in (Coccomini et al., 2022) and in (Abbas & Taeihagh, 2024, p. 1), "digital content, usually a video, audio, or image, that has been manipulated using deep learning algorithms to alter, replace, or superimpose the original content with new content that appears to be authentic". Today, deepfakes are widely spread across various realms such as politics, economics, and social issues, undermining both societal integrity and information security (Iufereva, 2023).

Recent advances in the field of GenAI during the last years have made deepfakes more realistic and harder to detect. The rapidly ongoing increase of AI-generated content is a growing social issue that poses serious challenges for contemporary society. The number of created images between Nicéphore Niépce’s first heliograph image and the appearance of the first digital camera has been estimated to be 15 billion analogue

images. With the frequent use of GenAI techniques for image creation today, the same number of 15 billion images was created in just one year after the release of the first accessible AI image generator in September 2022 (Attie, 2023; Tensetti & Rangsjö, 2024). Even though the importance of deepfake detection is widely recognized, the applied research remains limited. Deepfakes that are generated with the use of deep learning can also be detected with tools using deep learning. The study by Almars (2021) concluded that deep learning techniques showed good performance in deepfakes detection, but also that the quality of deepfakes is continuously increasing. Three years later, the study (Abbas & Taeihagh, 2024) also reported promising results, but with a similar comment that " the advances in deepfake generation methods are creating challenges concerning combating them" (p. 27).

Previous research has also examined both human capabilities and the effectiveness of AI tools in detecting deepfakes. Many of these studies (Groh et al. 2022, Somoray et al., 2024; Maiano et al., 2024; Moeßner & Adel, 2024) focused on synthetic datasets of AI-generated photos and videos, such as FaceForensics++, the Deepfake Detection Challenge (DFDC), SDXL, and Fooocus. One of the main limitations of such a method is that synthetic datasets fail to capture the complexity and contextual nature of non-synthetic deepfake occurring on social media. Furthermore, non-synthetic deepfake that have appeared on social media are rarely used in academic research, partly because such content is often removed due to moderation policies. Nevertheless, studies (Abbas & Taeihagh, 2024; Stroebel et al., 2023) emphasize the need for empirical analysis based on non-synthetic deepfake retrieved from social media to better assess the challenges of deepfake detection in real communicative environments.

The approach proposed in this paper compares human performance with AI-based tools in detecting deepfakes, focusing specifically on non-synthetic deepfakes that emerged during the Israeli–Palestinian war (2023–2024). The research questions that guided this study were:

RQ1: What visual errors, artefacts and inconsistencies can be identified in the selected deepfakes by human analysis?

RQ2: To what degree are GenAI based tools for deepfake detection correct and reliable?

2. Data Collection

The study was conducted as an empirical case-based analysis with AI-generated images gathered from multiple sources. Among these sources were established media agencies such as 'Reuters ', and the investigative platform 'Deutsche Welle' (DW). Moreover, the analysis involved educational materials used in the framework of the course 'Advanced Training Course in the Didactics of the Holocaust', given at the University of Florence (2024). The selected images appeared on social media during the Israeli-Hamas war in 2023-2024. The research sample consists of four images representing two different perspectives: two 'Pro-Palestinian' and two 'Pro-Israelian'. This study used the five freely available AI detection tools: 'Sightengine', 'Undetectable.ai', 'Decopy.ai', 'WasItAI', and 'Hive Moderation'. A choice that was motivated by several reasons. Firstly, the goal was to assess the degree of democratic accessibility of these tools for regular users — without payment and with user-friendliness. Secondly, the opportunity of immediate verification of digital content authenticity.

3. Results of Analysis

The results of the analysis are divided into two sections. The first section reports the findings of the human analysis, focusing on visual errors, artifacts, and inconsistencies identified in the selected deepfake images. The second section evaluates the performance of AI-based deepfake detection tools in terms of accuracy and reliability.

3.1 Human Analysis: Visual Errors, Artifacts, and Inconsistencies

The human analysis involved a detailed examination of the selected deepfake images to identify visual anomalies typical of AI-generated content. These included inconsistencies in environmental elements, unnatural features, and technical artefacts that reduce image authenticity.

Case 1. The dad with children in the Gaza Strip.

The example of a deepfake features an image of a father surrounded by his five children standing in front of a pile of rubble (See Figure 1). This photograph, widely disseminated on X and Instagram, has been viewed hundreds of thousands of times in connection with Israel's bombardment of Gaza. However, a community notice

on X now identifies the image as manipulated. "In the meantime, the image has been marked with a community notice, at least on X, that it is fake" (Eisele & Steinwehr, 2023). For instance, the man's right shoulder appears unnaturally elevated, disrupting anatomical proportion. Equally peculiar are the two limbs extending beneath his sweater, giving the illusion of unnatural emergence. Moreover, the hands of the two boys embracing their father's neck seem to merge indistinguishably, further distorting the composition. Additionally, discrepancies in the number of fingers and toes across several hands and feet in the image contribute to its overall inconsistency (see Figure 2).

Source: Personal archive of the authors; this example was presented by student Fausta Blasi during the Advanced Training Course in the Didactics of the Holocaust, University of Florence, 2024.

Figure 1: A father with children in the Gaza Strip (original photo)

Figure 2: Debanked AI-generated image showing typical inconsistencies and visual errors

Case 2. A Palestinian boy

The image depicts a child gazing upwards with his arm raised, surrounded by rubble and the bloodied face of a veiled woman, purportedly his mother, who is said to have died following a bombing in Gaza (see Figure 3). Initially shared by "several pro-Hamas accounts", the photograph quickly gained widespread circulation across numerous accounts on X, many of "which were flagged it as a fabricated creation" (Open, 2023). However, its authenticity has been questioned, with multiple users flagging the image as fabricated. Critical examination reveals significant inconsistencies: notably, the child's left hand is depicted with six fingers (see Figure 4). The presence of extra fingers, a common anomaly in AI-generated imagery, points to the likely use of free AI image generation software, which is prone to such errors. Additionally, a closer inspection of the child's eyes and eyelashes reveals disproportionate enlargement compared to the rest of the facial features.

Source: Personal archive of the authors; this example was presented by student Fausta Blasi during the Advanced Training Course in the Didactics of the Holocaust, University of Florence, 2024.

Figure 3: A Palestinian boy (original photo)

Figure 4: Debanked AI-generated image showing a boy with six fingers on his left hand

Case 3. Patriotic Illusion: Signs of AI Generation in Depicting the Israeli Army

The image depicts soldiers marching with Israeli flags through a settlement of bombed-out buildings, evoking a sense of patriotism (see Figure 5). However, upon closer inspection, several features undermine its authenticity.

The movement of the Israeli flags appears unnaturally uniform, the central street is improbably pristine, and the debris exhibits an unusual level of uniformity. Additionally, the destroyed buildings resemble one another closely, positioned at remarkably regular intervals. Collectively, the overall visual effect is excessively clean and lacks the irregularities characteristic of real-life scenes (see Figure 6). This type of artificial perfection, which often makes images look painted or staged, is a hallmark of AI-generated content.

Source: Eisele, I. and Steinwehr, U. (2023) Fact check: AI fakes in Israel's war against Hamas. DW, 10 November. Available at: https://www.dw.com/en/fact-check-ai-fakes-in-israels-war-against-hamas/a-67367744 [Accessed 19 July 2025].

Figure 5: March of the Israeli army (original photo)

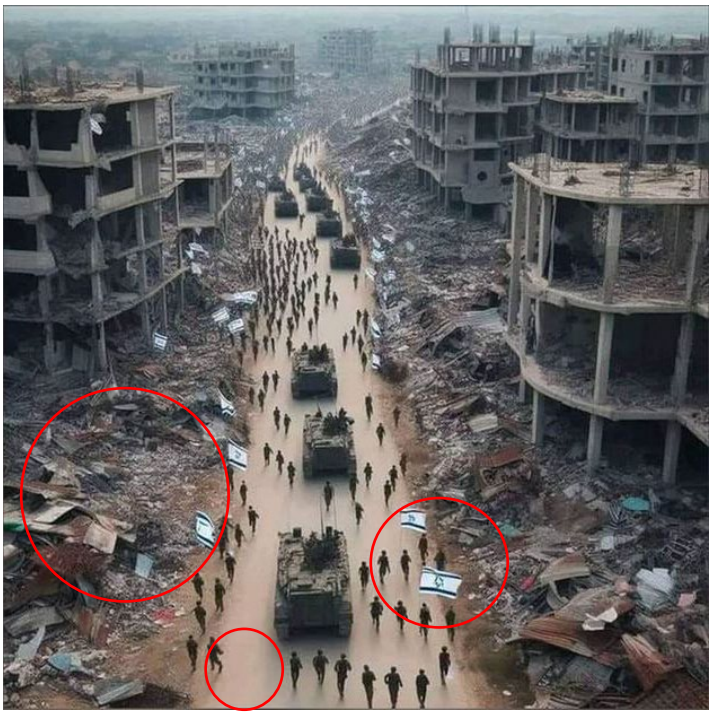

Figure 6: Debunked AI-generated image: lack of authenticity revealed through artificially flawless environment

Case 4. Crowds and Israeli soldiers

An image depicting crowds on balconies cheering and waving Israeli flags as soldiers march along a street was created using AI for inspirational purposes as well (see Figure 7). The image was widely shared on social media platforms accompanied by hashtags such as #HamasTerrorist and #IsraelFightsBack. On Facebook, it appeared with captions such as “The REAL Israel. Need we say more? AM YISRAEL CHAI!!” (The nation of Israel is alive) and its Spanish equivalent, emphasizing unity and resilience (Eisele & Steinwehr, 2023). However, the image has caused confusion among online users, many of whom question its authenticity and suspect it to be official propaganda.

The original post was uploaded on October 11 by Omri Shefi as part of a collection of thirteen AI-generated images depicting civilians and soldiers waving flags and celebrating. In his post, Shefi stated that the images were created with AI as part of psychological warfare, aiming to envision victory. His caption read: “Much of this warfare is psychological. So, let's imagine our victory picture, and so it will be!” (translated from Hebrew). Despite the clarification regarding its origins, the image raised skepticism among social media users, with some accusing it of being fabricated for political messaging. This example underscores the potential for AI-generated imagery to blur the lines between artistic expression and propaganda, especially in the context of geopolitical conflicts.

Source: Reuters Fact Check, 2023. Fact Check: Photo of cheering crowds waving Israeli flags at soldiers is AI-generated. Reuters, 30 October. Available at: https://www.reuters.com/fact-check/photo-cheering-crowds-waving-israeli-flags-soldiers-is-ai-generated-2023-10-30/ [Accessed 19 Jul. 2025].

Figure 7: Jubilant crowds waving Israeli flags at soldiers (AI-generated image)

Figure 8: Debunk: the image demonstrates typical AI errors and inaccuracies

3.2 AI-Based Detection Tools: Evaluation of Accuracy and Reliability

A set of free AI-based detection tools was used to analyze the selected images. Each image was uploaded for evaluation to the websites for the detection tools: 'Sightengine', 'Undetectable.ai', 'Decopy.ai', 'WasItAI', and 'Hive Moderation'. Since all images in the sample were AI-generated, it was expected that accurate tools would yield verification scores between 80% and 100%. Results within this range were marked with a "✅" indicating that the tool correctly identified the image as AI-generated and can be considered reliable. If the detection score was below 80%, it was marked with an "❌" signaling that the tool failed to accurately detect the image's artificial origin. The results of the analysis are presented below in Table 1.

Table 1: AI-detection tools results: overview of effectiveness

Detection tool	Picture 1 *'The dad with children in the Gaza Strip'*	Picture 2 *'A Palestinian boy'*	Picture 3 *'Jubilant Crowds Waving Israeli Flags at Soldiers'*	Picture 4 *'Crowds and Israeli soldiers'*
Sightengine	❌	❌	✅	❌
Undetectable.ai	❌	❌	✅	✅
Decopy.ai	❌	❌	❌	✅
WasItAI	✅	✅	✅	✅
Hive Moderation	❌	✅	✅	✅

Source: The authors' elaboration

4. Discussion

The study was divided into two main sections: 1) human deepfake analysis and 2) the evaluation of AI-based deepfakes detection. The first stage involved identifying and describing errors and inconsistencies in the selected images through human analysis. The detection in the first section relied on a critical examination of the images, assessing the logic of environmental elements, as well as recognizing common typical and earlier identified AI-generated errors and glitches such as humans with six fingers instead of five (Cramer, 2024; Jang et al., 2025). In the second stage, several free AI-based detection tools were employed to evaluate the authenticity of the selected AI-generated images. The results from this relatively small sample of images, showed that not all tools performed equally well and with surprisingly poor results. The most accurate was WasItAI, which correctly identified all images as AI-generated with high precision. Acceptable results also for the tool Hive Moderation

that only one failed in the detection of one image. The other tools that were tested: Sightengine, Undetectable.ai and Decopy.ai cannot be recommended for determining image authenticity, and must be classified as unreliable.

Unlike most prior studies (Groh et al. 2022, Somoray et al., 2024; Maiano et al., 2024), this research examines deepfakes in real media environments. The involved AI-generated images circulated in social media during the Israeli-Hamas war between the years of 2023 to 2024. By combining human analysis with automated detection tools applied to non-synthetic deepfake content shared during the Israel–Hamas war, this study provides grounded and timely insights to deepfake detection research. Authors see the four selected and analysed images as a sample representing the quality that AI-generated images had during 2023 to 2024. How to develop tools that could assess AI-generated image quality is an ongoing process (Aziz et al., 2025; Hartwig et al., 2025), but according to human assessment, the quality is continuously improving with the tool Midjourney as an example of high quality (Rangsjö &Tensetti, 2024; Cai et al, 2025). The humanly identified AI-artefacts, such as hands with four or six fingers, will probably not be there in the next generation of AI-generated images, and for dual reasons. Firstly, the next generation of image generation tools will have higher quality mentioned above. Secondly, the results from this study and others, and the general discussion on deepfakes, will probably lead to more detailed prompting where known glitches are hidden.

4.1 Limitations of the Study and Future Research

Some limitations of this study should be mentioned. First, the sample of AI tools was limited and included only free platforms, and it would be interesting to expand the sample in a broader comparison of tool accuracy and limitations. Second, the study focused solely on image deepfakes and did not cover other types of deepfake media, such as video, audio, music, text, or simpler AI manipulations like cheapfakes. Third, empirical material could preferably be extended beyond a narrowly focused military conflict to include other topics that are widely discussed on social media. Fourth, the detection tools that showed promising results should be further tested with larger data sets.

This paper could be seen as the first part of a longitudinal cross-sectional study covering deepfakes published from 2023 to 2024. This research field is a moving target, and a follow-up study on images generated during 2025 will probably generate different results. It would also be interesting to assess images from other contexts, but as concluded in Lundberg and Mozelius (2024), news media is the crucial area, while deepfakes in other areas could be accepted to a higher degree. Finally, it would be interesting to evaluate Google video generating tool Veo 3, which should be able to generate hands with five fingers (Kieffaber et al., 2025).

5. Conclusions

This research explored practical aspects of deepfakes detection. In a two-stage approach involving human analysis, and AI-detection tools analysis, the two research questions were answered. The human analysis identified the presence of extra fingers and toes, and disproportional limbs, eyes, and eyelashes in photos with humans. Regarding environments and buildings in the images, the human eye could identify structural repetitions and the artificial flawlessness and cleanness that give images a painted or staged appearance. Regarding the second research question, three out of five detection tools showed poor performance. The results were promising for one of the tools, but before any further conclusions or recommendations, the tool has to be further evaluated with more test data. Authors' recommendations are the same as brought up in the study by Janf et al. (2025), to label or watermark the AI involvement in image generation in a reliable way, and to increase deepfake literacy.

The research offers a conceptual basis for further empirical research regarding different types of disinformation and deepfakes. Finally, understanding the distinctive features of various disinformation types may serve as a basis for media literacy initiatives and for the development of the next generation of AI-based detection tools**.**

AI declaration: The describe detection tools were used in the the image evaluation and analysis, but all parts of this paper have been written without any use of AI.

References

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications, 252,* 124260. https://doi.org/10.1016/j.eswa.2024.124260

Ahmed, S., & Chua, H. W. (2023). Perception and deception: Exploring individual responses to deepfakes across different modalities. *Heliyon, 9*(10).

Almars, A. M. (2021). Deepfakes detection techniques using deep learning: A survey. *Journal of Computer and Communications, 9*(5), 20–35.

Attie, I. (2023, September 28). 19 Visual AI stats: AI-generated images in impressive numbers. *AI Secrets.* https://aisecrets.com/applications/visual-ai-stats/

Aziz, M., Rehman, U., Danish, M. U., & Grolinger, K. (2025). Global-local image perceptual score (GLIPS): Evaluating photorealistic quality of AI-generated images. *IEEE Transactions on Human-Machine Systems.*

Cai, Y., Zhou, Y., Yu, X., & Kim, W. (2025, May). Identifying usability challenges in text-to-image AI: A comprehensive comparison among mainstream platforms. In *International Conference on Human-Computer Interaction* (pp. 163–174). Springer Nature Switzerland.

Cramer, F. (2024). Speculative photography: An attempt (essay) of making visible (manifesto). *Artnodes, 34.*

Decopy.ai. (2025). *Decopy – AI detector and rewriter.* https://www.decopy.ai

Dickerman, L. (2000). Camera obscura: Socialist realism in the shadow of photography. *October, 93,* 139–153.

Eisele, I., & Steinwehr, U. (2023, November 10). Fact check: AI fakes in Israel's war against Hamas. *DW.* https://www.dw.com/en/fact-check-ai-fakes-in-israels-war-against-hamas/a-67367744

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences, 119*(1), e2110013119. https://doi.org/10.1073/pnas.2110013119

Hartwig, S., Engel, D., Sick, L., Kniesel, H., Payer, T., Poonam, P., Glockler, M., Bauerle, A., & Ropinski, T. (2025). A survey on quality metrics for text-to-image generation. *IEEE Transactions on Visualization and Computer Graphics.*

Hive Moderation. (2025). *Hive Moderation – AI-powered content moderation.* https://thehive.ai/content-moderation

Iufereva, A. (2023). Fake news as a distortion of media reality: Tell-truth strategy in the post-truth era. In *Proceedings of the 22nd European Conference on Cyber Warfare and Security, 22*(1), 610–615. 10.34190/eccws.22.1.1080.

Jang, E., Lee, H. M., Lee, S., Jung, Y., & Sundar, S. S. (2025). Too good to be false: How photorealism promotes susceptibility to misinformation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–8).

Kieffaber, J., Gandall, K., Foster, S. M., Jr., & McLaren, K. (2025). LLMs are bad judges. So use our classifier instead.

Lundberg, E., & Mozelius, P. (2024). The potential effects of deepfakes on news media and entertainment. *AI & Society,* 1–12.

Maiano, L., Benova, A., Papa, L., Stockner, M., Marchetti, M., Convertino, G., ... Amerini, I. (2024). Human versus machine: A comparative analysis in detecting artificial intelligence-generated images. *IEEE Security & Privacy, 22*(3), 77–86.

McQuire, S., Pfefferkorn, J., Sunde, E. K., Lury, C., & Palmer, D. (2024). Seeing photographically. *Media Theory, 8*(1), 1–18.

Moeßner, P., & Adel, H. (2024). Human vs. AI: A novel benchmark and a comparative study on the detection of generated images and the impact of prompts. *arXiv Preprint, arXiv:2412.09715.*

Open. (2023, November 4). La foto del bimbo palestinese con sei dita: «È generata con l'Intelligenza Artificiale». *Open.* https://www.open.online/2023/11/04/intelligenza-artificiale-bambino-palestina-foto-fake/

Palmquist, P. E., & Kailbourn, T. R. (2005). *Pioneer photographers from the Mississippi to the Continental Divide: A biographical dictionary, 1839–1865* (Vol. 2). Stanford University Press.

Pisch, A. (2014). *The personality cult of Stalin in Soviet posters, 1929–1953: Archetypes, inventions and fabrications.*

Reuters Fact Check. (2023, October 30). Fact check: Photo of cheering crowds waving Israeli flags at soldiers is AI-generated. *Reuters.* https://www.reuters.com/fact-check/photo-cheering-crowds-waving-israeli-flags-soldiers-is-ai-generated-2023-10-30/

Sightengine. (2025). *Sightengine – Image and text moderation API.* https://sightengine.com

Somoray, K., Miller, D., & Holmes, M. (2024). Human performance in deepfake detection: A systematic review. *SSRN.* https://ssrn.com/abstract=4955104

Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology, 7*(2), 83–113. https://doi.org/10.1080/23742917.2023.2192888

Tensetti, J., & Rangsjö, J. (2024). *Med AI som stöd: Lyft lärandet med AI* [With AI support: Lifting the learning with AI]. Liber AB.

Undetectable.ai. (2025). *Undetectable.ai – AI content humanizer.* https://undetectable.ai

WasItAI. (2025). *WasItAI – AI content detector.* https://wasitai.com