AI-Generated Content Detection: A Cat-and-Mouse Game of Technology and Ethics

Aadya Ajit Deshmukh[1], Pranavkumar A Bhadane[2]

[1]*Student, Artificial Intelligence and Machine Learning Department, Agnel Polytechnic*

[2]*Professor, Artificial Intelligence and Machine Learning Department, Agnel Polytechnic*

Abstract AI-generated content now appears across education, journalism, governance, and virtual environments. While it is great for efficiency and creativity, it also introduces risks of plagiarism, misinformation, and loss of trust. Detection methods have emerged to separate synthetic from human content, using statistical stylometry, linguistic and semantic analysis, image forensics, and large language models. But these systems are not perfect, as they show high false positive rates and limited adaptability to new generative models. This survey examines the technical basis of detection, its application in key domains, and the ethical concerns that accompany its application. Bias, privacy, and the danger of wrongful attribution are emphasized as critical challenges. The paper concludes with open problems and future directions for reliable and responsible detection.

Index Terms: Educational Integrity, AI-Generated Content Detection, Large Language Model, Text Authenticity Verification, AI-generated Images, News and Journalism, Convolutional Neural Networks, Diffusion Models, Synthetic Image Detection, Metaverse

I. INTRODUCTION

Generative AI systems can now produce text, images, video, and code with fluency and realism approaching human standards. Large language models generate coherent essays and dialogue; diffusion and adversarial networks synthesize high-quality images and video. These advances expand opportunities for education, communication, and media, but they also complicate questions of authorship, authenticity, and integrity.

Detection technologies have been developed to address these concerns. Stylometric approaches analyze word frequency and structure. Neural classifiers identify semantic irregularities or contextual inconsistencies. Image forensics targets artifacts introduced by GANs or diffusion models, while hybrid systems combine text and image pipelines. Detectors are unreliable in practice, despite their progress. False positives and errors can harm students accused of plagiarism, undermine journalists reporting authentic stories, or wrongly flag legitimate online content.

The aim of this survey is to assess AI content detection not only as a technical challenge but also as a social and ethical problem. It highlights the difficulty of balancing effective detection with fairness and trust. Four application areas are examined: education, journalism, the metaverse, and governance. For each, technical detection methods are reviewed alongside the ethical risks. Broader cross-cutting issues such as bias, privacy, and accountability are also discussed.

II. LANDSCAPE OF AI GENERATED CONTENT

Recent advances in generative models have transformed the creation process. Models such as Generative Adversarial Networks (GANs), diffusion models, and large language models (LLMs) produce text, images, audio, and video that are often on par with human work in fluency and realism. These systems are trained on massive datasets and learn statistical patterns that allow them to generate coherent and contextually appropriate outputs.

Text generation: LLMs like GPT-style architectures produce essays, reports, code, and dialogue. They capture semantic relationships across long contexts and adapt easily to user prompts. While they are powerful, they can fabricate facts, mimic bias, and imitate individual writing styles, making human detection unreliable.

Image and video generation: GANs and diffusion models can synthesize highly realistic faces, environments, and moving sequences. Visual outputs often contain subtle artifacts such as frequency irregularities, inconsistencies in shading, or structural distortions, etc., that specialized detectors attempt to identify. Newer models reduce these artifacts, narrowing the detection margin.

Multimodal generation: Systems that integrate text and vision, such as text-to-image pipelines, enable cross-domain synthesis. They generate images from captions, captions from images, or videos from scripts. Multimodality complicates detection because artifacts differ across modalities and can reinforce one another.

Accessibility: Tools are widely available through APIs and consumer applications. This reduces barriers, and thus, malicious use such as misinformation campaigns, academic dishonesty, or fabricated evidence becomes easier. As quality improves, manual inspection alone is insufficient for verification.

The rapid technical progress of generative systems has made detection both necessary and difficult. Each new model family changes the characteristics that detectors must target. This dynamic creates a persistent gap between generation and detection methods. The models and detectors must remain in a constant cat-and-mouse game.

III. TECHNICAL FOUNDATION OF AI CONTENT DETECTION

AI content detection methods can be grouped into three main categories:

1. text-based
2. image- and video-based, and
3. hybrid systems,

which rely on different signals to separate human- and machine-generated content.

A. Text-based Detection

Early methods looked at simple features such as word choice, sentence length, and writing style. Classifiers like support vector machines used these patterns to separate human and AI writing. More recent detectors use large language models (LLMs). These analyze coherence, token probabilities, and semantic flow to judge whether a passage is machine-written. They work well on standard datasets but often fail when text is paraphrased, slightly edited, or written in another language.

B. Image and Video Detection

Visual detectors search for traces left by generative models. GAN images may show unusual textures or frequency patterns, while diffusion models leave subtler statistical artifacts. Convolutional neural networks (CNNs) trained on large datasets are commonly used for this task. To keep up with new generators, some systems adopt incremental learning, updating their knowledge without retraining from scratch. Detectors also track consistency across frames in videos, since generative models sometimes introduce irregular movements or lighting.

C. Hybrid and Multimodal Detection

Some detectors combine methods across different content types. A system might check both the text and images in a news article, then merge the results into a single decision. This gives broader coverage, especially for content that mixes media, but also inherits the weaknesses of each individual detector.

D. Limitations

Despite progress, detection tools remain unreliable. AI content can be masked by simply making small edits, while human-made content can be wrongly flagged as AI-generated. Detectors trained on one generation model can fail against newer ones. These mistakes can have serious consequences for students, journalists, and others affected by detection systems. These weaknesses show that technical performance and ethical concerns cannot be separated.

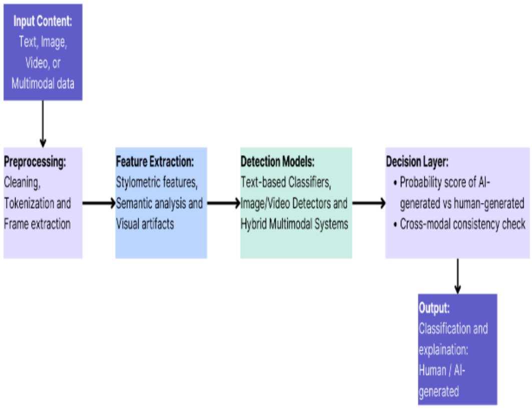
Input Content:
Text, Image, Video, or Multimodal data
Preprocessing:
Cleaning, Tokenization and Frame extraction
Feature Extraction:
Stylometric features, Semantic analysis and Visual artifacts
Detection Models:
Text-based Classifiers, Image/Video Detectors and Hybrid Multimodal Systems
Decision Layer:
• Probability score of AI-generated vs human-generated
• Cross-modal consistency check
Output:
Classification and explaination: Human / AI-generated

Fig 1 – Workflow of AI-generated content detection

IV. APPLICATIONS AND ETHICAL CONCERNS

AI content detection is crucial in areas where trust and authenticity matter. Four areas show the strongest need: education, journalism, virtual environments, and governance. In each area, detection methods can be helpful but also raise ethical issues.

A. Education

Students now use AI tools to write coursework, including essays, assignments, and lab reports. AI detectors can help teachers spot AI-written work and ensure that students are actually learning. However, a student might be wrongly accused if their writing style resembles machine output, while actual misuse might go undetected with minor edits. These situations damage trust between students and institutions. The ethical issue is not just about catching dishonesty but also about ensuring that innocent students are not punished.

B. Journalism and Media

Fake images and false stories threaten the integrity of the news we believe. Deepfakes of important people can easily deceive the public. Detection methods based on CNNs can reveal signs of manipulation, but they are not perfect. Mislabelling a genuine photograph as fake can hurt a journalist's reputation. Inaccurate results can significantly affect both journalists and the public. The ethical challenge is balancing accuracy with the right to report freely while avoiding unfair censorship.

C. The Metaverse and Virtual Worlds

In immersive environments, AI can generate avatars, virtual spaces, and dialogue that blur the line between human and machine. This creates chances for creative interaction but also raises risks of impersonation, scams, and identity theft. Detection methods can help prevent this, but they also bring up concerns about privacy. Monitoring user activity may protect some individuals but intrude on others. The ethical tension is keeping virtual spaces safe without intrusive surveillance.

D. Governance and Security

AI-generated content is dangerous in politics, cyberattacks, and online misinformation. Detectors that analyze language patterns and context can help identify suspicious material. False positives can silence the truth, while false negatives can allow harmful propaganda to spread. Governments and companies that control detection systems also face accountability questions. Who decides what is labeled as "AI-generated," and who is responsible when mistakes happen?

Table 1 - Applications vs. Ethical Concerns

Domain	Detection uses	Ethical Risks	References
Education	Spotting AI-written essays, assignments, and lab reports	False accusations of plagiarism, privacy intrusion	[5] Liu et al. (2024); [2] Murugesan (2023)
Journalism	Identifying manipulated images and deepfake news	Mislabeling genuine content, censorship risks, reputational harm for journalists	[4] Jagadish and G. J. S. (2024); [7] Zhou (2023)
Metaverse / Virtual Worlds	Detecting fake avatars, dialogue, and immersive spaces	Identity theft, impersonation, intrusive surveillance of users	[1] Basyoni and Qadir (2023); [14] Sanderson et al. (2023)
Governance & Security	Filtering misinformation, political deepfakes, and propaganda	Over-censorship, accountability gaps, bias in detection systems	[3] Gowrishankar et al. (2024); [16] Mökander et al. (2023)

V. Case Study

EduGuard-LLM was designed in response to the increasing use of generative AI tools such as ChatGPT and Copilot by students to complete assignments. This trend makes it difficult for teachers to evaluate a student's skills and undermines academic integrity. The model leverages the text analysis power of LLMs to detect hidden patterns that distinguish AI-written text from authentic student writing.

Data and Training:

The developers trained EduGuard-LLM on twenty-one publicly available datasets containing more than 164,000 text samples. These included elementary school essays, college entrance exam papers, and lab reports, as well as outputs from mainstream AI systems. Seventeen datasets were used for training, and four external validation sets represented primary, middle, high school, and university writing.

Performance and Application:
EduGuard-LLM achieved strong results, with an accuracy of 93.96 percent on the training set and between 93.97 and 95.01 percent on the four validation sets. These consistent results show that the model can detect AI content at multiple educational levels. In practice, when collections of essays were analyzed, the system was able to flag a significant number of AI-generated submissions, and most of these results were confirmed by reviewers. For example, in descriptive writing material, it successfully identified linguistic markers in AI-written essays that human readers overlooked.

Limitations and Ethical Issues:
EduGuard-LLM faces limitations despite being highly accurate. Its performance declines when students paraphrase or lightly edit AI-generated text, and it is less effective with non-English material because of dataset bias. False positives remain a critical concern since wrongly flagging student work can lead to disputes or unfair penalties. The authors emphasize that the tool should be treated as a decision-support system, not a final authority, and that the final result should always involve human review.

VI. ETHICAL AND SOCIAL CHALLENGES

EduGuard-LLM shows how LLM-based detectors can be used as technical support for safeguarding academic integrity. Its high accuracy across multiple educational stages shows promise for real-world use, but its deployment must be guided by clear institutional policies and educator involvement to ensure fairness and maintain trust in education.

VI. ETHICAL AND SOCIAL CHALLENGES

Across these domains, the same problems appear again and again. Detection is not only a technical task but also a social and ethical one.
Bias in detection: Detectors are trained on specific datasets. If those datasets are skewed, results may unfairly target certain writing styles, languages, or cultural expressions, which results in false positives. This creates unfair results and risks discrimination.

Privacy and surveillance: Effective detection often requires scanning large amounts of user data. In education or online platforms, this can feel intrusive. The right to privacy is just as important as verifying authenticity.

False positives and reputational harm: When a detector wrongly flags human work as AI-generated, the consequences can be serious. For ex., students accused of cheating, journalists accused of spreading fake news, or individuals losing credibility. The harm from a single mistake can outweigh the benefits of detection.

Control and accountability: Detection tools are usually managed by large institutions, governments, or corporations. This raises questions about transparency and fairness. Who decides what is considered AI-generated? Who is responsible when the system gets it wrong? Detection itself could become a tool for censorship or control for large organizations.

Legal and ownership disputes: Questions remain about the authorship of AI-generated work. Copyright, intellectual property, and accountability laws lag behind technology, leaving uncertainty for creators and users alike. This is especially true and common on social media.

VII. FUTURE RESEARCH DIRECTIONS

AI content detection will need to advance in several ways to keep up with new generative models.
Extensible detection: Current detectors often fail when facing unfamiliar models. Future work should focus on systems that can update efficiently as new generators appear, without needing full retraining. Incremental learning and adapter-based methods are promising in this area.

Multimodal detection: Many applications involve content that mixes text, images, audio, or video. Stronger systems will need to check multiple modalities together and cross-verify signals, rather than treating them separately.
Human-in-the-loop approaches: Fully automated detection can lead to many mistakes. Combining machine classifiers with human review, especially in

high-stakes areas like education or journalism, may reduce errors and improve fairness.

Policy and governance: Rules are needed to ensure transparency and accountability alongside AI detectors. Clear processes are needed to handle disputes, correct errors, and prevent misuse of detectors for censorship.

Ethical safeguards: More research needs to be done on how to reduce bias, protect privacy, and minimize reputational harm. For example, detectors could provide probability scores instead of absolute judgments, leaving space for human interpretation.
These directions highlight that detection is not only a technical race against generative models but also a matter of building fair, accountable systems that serve the public.

VIII. CONCLUSION

This survey highlighted how detection works, where it is applied, and why it matters. The main lesson is that no detector is perfect, and its mistakes carry real costs. AI-generated content is now a common part of education, journalism, virtual environments, and governance. It brings efficiency and creativity but also risks trust and integrity in these areas. Detection methods offer partial solutions, yet all remain limited. Protecting fairness in schools, accuracy in media, safety in virtual spaces, and credibility in governance depends on more than algorithms alone. Errors, bias, and lack of adaptability mean that technical progress cannot be separated from ethical responsibility. For AI to play a constructive role in a society, technical innovation must be matched with ethical safeguards and clear policies. Future systems must balance accuracy with transparency, privacy, and accountability.

REFERENCES

[1] L. Basyoni and J. Qadir, *AI Generated Content in the Metaverse: Risks and Mitigation Strategies*, IEEE, 2023.

[2] S. Murugesan, “The Rise of Ethical Concerns about AI Content Creation: A Call to Action,” *IEEE IT Professional*, Apr. 2023.

[3] G. J. Gowrishankar, S. Deepak, and M. Srivastava, *Countering the Rise of AI-Generated Content with Innovative Detection Strategies and Large Language Models*, IEEE, 2024.

[4] T. Jagadish and G. J. S., *Detection of AI-Generated Image Content in News and Journalism*, IEEE, 2024.

[5] L. Liu, D. Zhang, B. Yan, and D. Wu, *EduGuard-LLM: An AI-Generated Content Detector Using Large Language Models for Safeguarding Educational Integrity*, IEEE, 2024.

[6] H. Zhang, *Integrating Multicore SVM With Enhanced Residual Networks for AI Content Recognition*, IEEE, 2024.

[7] E. P. Zhou, *The Fallibility of AI Content Detectors*, IEEE, 2023.

[8] S. Tang, P. He, H. Li, W. Wang, X. Jiang, and Y. Zhao, “Towards Extensible Detection of AI-Generated Images via Content-Agnostic Adapter-Based Category-Aware Incremental Learning,” *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 2883–2896, 2025.

[9] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu, “Fusing Global and Local Features for Generalized AI-Synthesized Image Detection,” *arXiv preprint*, Mar. 2022.

[10] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On the Detection of Synthetic Images Generated by Diffusion Models,” *arXiv preprint*, Nov. 2022.

[11] Y. Quan and C.-T. Li, “Improving Detection of Unprecedented Anti-forensics Attacks on Sensor Pattern Noises Through Generative Adversarial Networks,” in *Proc. ICPR 2022 Workshops*, Lecture Notes in Computer Science, vol. 13646, 2023.

[12] Q. Xu, X. Jiang, T. Sun, H. Wang, L. Meng, and H. Yan, “Detecting Artificial Intelligence-Generated Images via Deep Trace Representations and Interactive Feature Fusion,” *Information Fusion*, vol. 112, 2024.

[13] T. Say, M. Alkan, and A. Koçak, “Advancing GAN Deepfake Detection: Mixed Datasets and Comprehensive Artifact Analysis,” *Applied Sciences*, vol. 15, no. 2, Jan. 2025.

[14] C. Sanderson, D. Douglas, and Q. Lu, “Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects,” *arXiv preprint*, Apr. 2023.

[15] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," *arXiv preprint*, 2023.

[16] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: a three-layered approach," *arXiv preprint*, 2023.

[17] A. Knott, D. Pedreschi, R. Chatila, T. Chakraborti, S. Leavy, R. Baeza-Yates, D. Eyers, A. Trotman, P. D. Teal, P. Biecek, S. Russell, and Y. Bengio, "Generative AI models should include detection mechanisms as a condition for public release," *Ethics and Information Technology*, 2023.

[18] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity," *arXiv preprint*, Jan. 2023.

[19] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M. B. Amin, and B. Kang, "Survey on Explainable AI: Approaches, Limitations and Applications Aspects," *Human-Centric Intelligent Systems*, 2023.