INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 1, PP. 1-9
Manuscript received January 12, 2026; revised January, 2026. doi: 10.24425/ijet.2026.157883

Artificial Intelligence in Education and Information Ecosystems: Learning Outcomes, Virality and the Social Impact of Automation

Miłosz W. Romaniuk, Andrzej Manujło, and Roman Androszczuk

***Abstract*—AI is no longer just a tool. It has become a fundamental part of our educational and information ecosystems. This study investigates how AI-generated assignments and content algorithms are changing the way students and scholars interact with knowledge. While the efficiency gains are obvious, our findings point to a deeper problem: a growing cognitive dependency that could weaken critical thinking. By connecting educational technology with information studies, we provide a roadmap for updating academic literacy and curriculum design for the AI era.**

***Keywords*—artificial intelligence; higher education; learning outcomes; information ecosystems; content virality**

Introduction

AI has evolved from a discrete classroom tool into a fundamental piece of our educational and information infrastructure. It is now a force that dictates how knowledge is produced and, more importantly, how it circulates. This perspective expands upon our previous analyses of how digital technologies are integrated into higher education [1] and how PhD students adapt to the digital age [2]. While we have previously documented the benefits of specific ICT tools for enhancing research practices [3], we must now address the systemic impact of widespread automation.

Interestingly, research in this area is often siloed. This article takes a cross-domain perspective to show how automated task generation might boost short-term engagement while fostering a dangerous cognitive dependency. At the same time, we look at how AI-driven "virality models" can privilege attention-grabbing content over actual scientific quality. We tackle three key questions. How does AI-assisted learning affect long-term student engagement? How do algorithms decide what academic content stays visible? And how do these forces combined change the way we define academic literacy? By merging these perspectives, we aim to provide a roadmap for a curriculum, and a policy, that can survive in an automated information ecosystem, ensuring that the digital competencies we previously identified as essential continue to support, rather than replace, critical human engagement.

I. How Does the Use of Artificial Intelligence in Automatic Task Generation Impact Student Learning Outcomes in Computer Programming Education?

AI applications in education, especially in computer science, have recently been attracting significant interest because of the potentially high degree of automation and content personalization. Tools such as ChatGPT, OpenAI Codex, and custom systems like PyTaskSyn enable new ways for automatically generating programming tasks tailored to the individual needs of learners. The first studies indicate that AI-driven task generation may increase student motivation, self-efficacy, and engagement at least during introductory programming courses. However, long-term evidence about the effectiveness of AI-generated tasks remains scarce, particularly regarding knowledge retention, development of critical thinking skills, and task reliability. [4] [5]

This paper systematically reviews existing research on AI-generated programming tasks. By analyzing methodological approaches, empirical findings, and ethical considerations, we aim to assess the educational impact of these AI tools. Furthermore, we propose a design framework for integrating narrative-driven learning elements, such as the 'Mr. Square' case study, into AI-enhanced educational systems, with the objective of improving student engagement and problem-solving capabilities [6] [7]. Inclusion criteria included peer-reviewed studies, conference papers, and systematic reviews on AI-driven task generation, adaptive learning systems, or automated assessment in computer programming education that were published between 2018–2025. Publications were eligible if they reported empirical results, described the methodological background, or discussed pedagogical and ethical aspects regarding AI-supported task design. This excluded papers that focused on AI-assisted code generation rather than task generation, general discussions about AI in education without any programming context, purely theoretical discussions without a core empirical backing, and works conducted outside computer science education. Duplicate reports and non-scholarly sources, such as blogs and opinion pieces, were excluded.

Authors are with The Maria Grzegorzewska University (APS), Warsaw, Poland (e-mail of corresponding author: mromaniuk@aps.edu.pl). Abstract, introduction, and conclusions are written by M. W. R., as is the overall edition of the paper. Chapter I is written by A. M..; chapter II by R. A;

A. Theoretical Foundations of AI-Generated Task Design

AI systems are designed to dynamically generate tasks based on learners' prior performance, preferences, and learning trajectories. This personalized approach to each user is consistent with the adaptive learning framework, which emphasizes tailored instruction to meet the diverse needs of students. AI-driven systems can adjust task difficulty in real time, offering more precise learning opportunities [8] [9]. AI-generated tasks support constructivist learning theories, where students actively construct knowledge through problem-solving and exploration. By providing iterative task sets and immediate feedback, AI fosters environments conducive to generative pedagogies, where students learn through inquiry, experimentation, and application [10] [11]. By automatically generating tasks at different levels of complexity, an AI system can potentially manage cognitive loads for novice learners by reducing extraneous cognitive load. Yet scaling such scaffolding for complex, multi-step problems remain a challenge as AI often fails at the task of generating tasks which seamlessly integrate several programming concepts. [5] [6]. Gamification theories support embedding game design principles into the generation of AI-driven tasks. According to theories of narrative learning, stories activate learners' emotions, therefore attracting and maintaining students' motivation and persistence. Results from analyzed papers indicate that incorporating AI-generated tasks into a broader narrative structure enhances both engagement and learning retention. [4] [6].

B. Literature Review: Methodology and Evaluation of Empirical Studies

Empirical research on AI-generated programming tasks has predominantly employed quasi-experimental designs, with varying levels of methodological rigor. Most studies have focused on undergraduate students, with a few extending to primary and secondary education contexts. Sample sizes have ranged from small (n=20) to large (n=230), and outcomes have often been assessed using a combination of self-report measures, performance-based assessments, and computational thinking tests [12]. Sample sizes: The number of participants varied widely across studies:

- Small-scale studies, with less than 50 participants, often focused on more controlled, qualitative observations or pilot tests. For instance, Jacobs et al. (2025) included in their work n=45 participants, while Binhammad et al. (2024) used in their research n=30 students. [8] [9].
- Medium-scale studies ranged between 50–150 participants, thus often allowing for more robust statistical analysis. Notably, Yilmaz & Karaoglan Yilmaz (2023) included n=100 undergraduates in their quasi-experimental design [5].
- Large-scale studies like Zhang & Li, 2024, with n=230 students and Sie & Lin, 2025, also with n=230 participants, provided valuable insights into the impact AI has on different educational levels and were better equipped for generalizing findings [9].

Study Methodology: Predominantly quasi-experimental designs with pre/post tests, experimental groups, and control groups, though randomized controlled trials (RCTs) remain rare. Quantitative data are often supplemented with qualitative insights, including learner surveys and feedback [12].

C. Synthesis of Results: Task Generation and Learning Outcomes

1) Strengths of AI-Generated Tasks

AI-generated exercises lead reliably to increased student motivation and engagement, at least for introductory types and routine problems. Problems generated via AI also foster an active, creative approach to problem solving [8] [9].

2) Limitations and Gaps

AI-generated tasks tend to struggle with more complex problems that require multi-step reasoning and the integration of various programming concepts. For advanced learners, tasks that involve high-level abstraction remain challenging [9] [12]. While AI tools enhance short-term learning outcomes, evidence on their impact on long-term retention is inconclusive [4] [8]. AI assistance may inadvertently reduce students' capacity to solve problems independently, leading to overreliance [6] [12].

3) Moderating Factors

More advanced learners profit less from automatically generated tasks than novices, who require more scaffolding. Such findings indicate a more focused role for AI tools on novice learners [10] [11]. Tasks generated within structured pedagogies such as Project-Based Learning (PBL) or flipped classrooms yield better results, aligning AI tools with a more active an learning environment [4] [5].

D. Ethical Considerations and Risks in AI Task Generation

Ethical Considerations and Risks in AI Task Generation Excessive reliance on the support of AI tools undermines critical thinking and fosters metacognitive strategies. Since learners depend on AI for task generation, their problem-solving processes will be less autonomous [9] [10]. AI-generated tasks may contain errors or inconsistencies—known as 'hallucinations'—which could negatively affect the learning process by introducing inaccurate or misleading content [6] [12]. The use of AI-driven platforms often involves collecting sensitive learner data, raising concerns about data privacy, storage, and potential misuse under regulations such as GDPR [11] [8]. If the underlying models are based on non-representative datasets, AI systems themselves bear the danger of introducing bias; this might be in the form of difficulty or via cultural assumptions. This involves a risk of inequity, particularly with students from diverse backgrounds [9] [10].

E. Implementation in Educational Systems: Moodle and CodeRunner

Moodle, in conjunction with the CodeRunner plugin, provides an ideal platform for integrating AI-generated tasks. CodeRunner automates task evaluation, which allows for real-time feedback, especially useful for large cohorts and programming assignments in languages like Python and C++ [8] [12]. AI tools can generate an extensive range of tasks automatically, adapting to student progress and providing scalable learning opportunities. CodeRunner enhances the learning process by providing immediate feedback on students' submissions, facilitating faster mastery of concepts [9] [11].

F. *Case Study: Mr. Square - A Narrative-Driven Learning Framework*

Mr. Square serves as a gamified, narrative-driven design to improve student engagement. By embedding programming tasks within a storyline, students are not only solving problems but also contributing to a character's journey. This approach leverages the power of storytelling to maintain motivation and ensure that tasks feel meaningful [5] [12]. AI generates coding challenges tied to the evolving story of Mr. Square. Students interact with these challenges while receiving feedback via CodeRunner. As students progress, they unlock subsequent stages of the narrative, providing a sense of accomplishment and advancing learning objectives. This narrative structure encourages active participation and persistence, fostering a deeper connection with learning tasks. The integration of gamification also aligns with motivational theories such as ARCS (Attention, Relevance, Confidence, Satisfaction) [6] [10].

G. *Future Directions and Research Gaps*

Future AI tools should aim to generate tasks that involve complex problem-solving, combining multiple concepts and requiring critical thinking, as opposed to solely focusing on individual concepts [9] [10]. Long-term studies are needed to assess how AI-generated tasks influence retention, independence, and the transferability of skills [11] [8]. Expanding the capabilities of AI to facilitate collaborative task generation would provide students with opportunities to work on team-based problems, better simulating real-world programming scenarios [5] [12].

H. *Conclusions*

AI-generated programming tasks have the potential to significantly improve computational thinking, motivation, and engagement, especially simpler, single-concept problems. Creating tasks that require higher-order problem-solving and guaranteeing long-term retention, however, continue to present difficulties. Integrating AI-generated tasks into pedagogical frameworks like project-based learning and flipped classrooms is essential to optimizing the advantages of AI-driven educational tools. While previous research indicates several possible advantages of AI-generated programming tasks, including improved motivation, engagement, and short-term performance, these results should be interpreted cautiously. All reported advantages remain provisional and cannot be considered robust until validated through large-scale, longitudinal research and rigorous randomized controlled trials (RCTs). Current evidence is largely based on small or medium cohorts, short intervention periods, and quasi-experimental designs, which limits the reliability, generalizability, and long-term predictive value of the results. Consequently, any claims regarding the sustained educational impact of AI-driven task generation should be treated as temporary, preliminary, and subject to future verification. [4] [6].

II. Artificial Intelligence in Predicting Online Content Virality: Opportunities and Threats in the Fight Against Disinformation

A. *Introduction*

In the age of digital transformation, artificial intelligence (AI) plays a key role in shaping the dynamics of online information dissemination. Predicting content virality - the ability of online materials to spread rapidly and massively is a key application area for AI, characterized by a dual nature. On the one hand, machine learning-based tools enable early detection of potentially viral content, helping identify false narratives before they reach a mass scale. [13] On the other hand, these same technologies can be used to create and amplify disinformation, posing serious threats to democratic societies, and the stability of electoral processes. The complexity of this phenomenon stems from the multidimensional nature of virality, which depends on social platform algorithms, user behavior, and sociopolitical context. Predictive AI models process big data sets to identify features that determine virality, analyzing both the content of messages and the dynamics of their spread across social networks [14]. Of particular importance in this context is the analysis of emotions evoked by content and user reactions, which allows for the prediction of their viral potential [15]. Disinformation, defined as intentionally false or misleading information, achieves high virality thanks to strong emotions such as fear or anger. [16] The development of generative AI further complicates the situation by creating synthetic content that is difficult to distinguish from authentic content, requiring new methodological and regulatory approaches. This chapter examines the opportunities and risks associated with using AI to predict content virality and combat disinformation, considering both the technology's potential and its technical, ethical, and social limitations.

B. *Purpose and methodology of the review*

The aim of this chapter is to systematically analyze the role of artificial intelligence in predicting the virality of online content and assess its potential and limitations in combating disinformation. The study focuses on identifying the technological mechanisms used to detect and amplify viral content, assessing the effectiveness of various methodological approaches, and analyzing the ethical and social implications of using AI in this area. This review was conducted according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The following databases were searched: Web of Science, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar. Keyword combinations in English were used: "artificial intelligence," "machine learning," "viral content prediction," "virality," "misinformation detection," "fake news," "disinformation," "social media," "content moderation," "deepfakes," and "fact-checking."

C. *Theoretical foundations of viral content in the digital environment*

The virality of online content is a complex phenomenon determined by the interaction of technological, psychological, and social factors. According to Berger and Milkman's model, the virality of content depends on its ability to evoke high-arousal emotions (arousal), such as admiration, anger, anxiety, or amusement. [17] Content that evokes strong emotional responses is more likely to be shared than neutral content or content that evokes low-arousal emotions (sadness, happiness). Social media platform algorithms play a key role in shaping the trajectory of content spread. Machine learning-based recommendation systems analyze metadata - the number of shares, likes, comments, and time spent on content to predict and simultaneously influence the trajectory of information spread. [18] These mechanisms are optimized to maximize user engagement, which often leads to a preference for controversial and emotionally intense content.

Empirical studies indicate that false information spreads faster and more widely than true information. Vosoughi and colleagues found that fake news on Twitter was 70% more likely to be retweeted than true information and took about six times less time to reach 1,500 users. [18]

D. Applications of artificial intelligence in predicting content virality

Modern virality forecasting systems utilize advanced machine learning architectures, including models based on natural language processing (NLP), neural networks, and ensemble learning. These models can be classified according to several key technological dimensions. Early-stage prediction models analyze content characteristics before publication or in the initial stages of dissemination. They utilize sentiment analysis, linguistic complexity, the presence of specific keywords, and narrative structure. Bandeli et al. demonstrated that models based on Random Forest and Gradient Boosting algorithms achieve 78–82% accuracy in predicting the virality of news articles based on analysis of the title, lead, and first paragraphs of the text [14]. Temporal cascade models account for the dynamics of content spread over time by analyzing the growth trajectories of shares, likes, and comments. These approaches often use neural networks such as LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Units) to model temporal sequences. Studies show that incorporating temporal data increases prediction accuracy by 15–20 percentage points compared to models that analyze only static content features. Multimodal models combine text, image, and video analysis, leveraging deep learning architectures such as CNNs (Convolutional Neural Networks) for image processing and Transformer-based models (BERT, GPT, RoBERTa) for text analysis. Gondwe (2025) demonstrated that BERT- and GPT-3-based models achieve F1 scores of 0.89–0.92 for real-time fake content classification by simultaneously analyzing semantic context and sentiment of utterances [19]

AI systems identify multidimensional content characteristics that correlate with high virality. Krstovski et al. (2024) developed the Evons dataset, which contains over 6,000 news articles (both true and false) and their virality data. [15] Analysis of this dataset revealed that the key predictors of virality are:

- Linguistic features: polarity of sentiment (especially strongly negative or positive), syntactic complexity (paradoxically, simpler structures achieve higher virality), presence of words evoking strong emotions (so-called arousal words).
- Network features: early adoption by nodes with high centrality in the social network, clustering of shares in specific topic groups, bridging between different communities.
- Temporal features: the rate of growth of engagement in the first hours of publication, the occurrence of a "second wave" of engagement after a period of stagnation, consistency with the current news cycle.

Metadata features: source credibility, author's publication history, thematic relevance to current trends. AI enables early detection and mitigation of the spread of disinformation through several technological mechanisms. Predictive models classify content as potentially viral and false based on analysis of language, images, and external links, allowing platforms to flag or reduce its reach before it reaches mass scale. Integrating automated fact-checking techniques with social network analysis increases the effectiveness of content moderation. Hassan et al. (2019) developed the ClaimBuster system, which uses NLP to automatically identify verifiable claims in text content, achieving a precision of 0.85 and a recall of 0.78 in identifying claims requiring fact-checking [20]. This system, integrated with databases of verified facts, allows for semi-automatic content verification in near real time. The VERA.ai project is an example of integrating AI with expert crowdsourcing to detect and verify false information, including deepfakes [21]. The system combines automated content analysis with expert reviewers, achieving a balance between the scalability of automation and the accuracy of human judgment. Studies show that hybrid systems achieve 12-18% higher accuracy than fully automated systems.

Personalization of recommendations that prioritize verified and educational content is a potential tool to counteract the filter bubble effect. AI models predict the virality of educational content, supporting public health and election information campaigns. During the COVID-19 pandemic, social media platforms implemented algorithms promoting content from official medical sources, which, according to Renda and Simonelli (2025), contributed to a reduction in exposure to medical misinformation by approximately 34% among users actively seeking information about the pandemic [22]. However, paradoxically, these same personalization mechanisms can reinforce polarization. Algorithms that optimize engagement can steer users toward content increasingly consistent with their existing beliefs, even if that content is verified. This challenge requires the development of diverse aware algorithms that balance personalization with exposure to diverse perspectives.

E. Artificial intelligence as a tool for amplifying disinformation

The development of large language models (LLMs) and generative multimodal models has dramatically lowered the barriers to entry for producing persuasive disinformation. Models such as GPT-4, Claude, and Gemini can generate politically relevant false content indistinguishable from real news, amplifying disinformation narratives [23]. Deepfakes - synthetic video or audio content using generative adversarial networks (GANs) and diffusion models represent a particularly dangerous category of disinformation. These technologies enable digital impersonation and the creation of realistic statements by politicians, celebrities, or ordinary citizens that never actually occurred [24]. The 2024 election campaigns saw cases of AI-generated disinformation, false images and audio recordings spread rapidly thanks to social media algorithms, targeting specific audiences [25]. The perceived credibility of deepfakes currently reaches around 75-85% for average viewers, meaning most people cannot distinguish synthetic content from authentic content without specialized detection tools. This problem is particularly relevant in the context of the so-called "liar dividend "an effect in which the mere awareness of deepfake technology allows politicians and public figures to deny authentic, compromising material by claiming it is fake.

AI-powered social bots pose another dimension of the threat. Cresci and Ferrara (2025) demonstrated that modern bots using language models can simulate human behavior at a level that makes detection difficult even for advanced systems [24]. In election campaigns, bots amplifying false content achieve billions of views, creating the illusion of mass support for specific narratives. Analysis of data from the 2020 US presidential election revealed that approximately 15-20% of accounts actively spreading political disinformation were bots or semi-automated accounts. These accounts accounted for approximately 45% of the total number of disinformation content shares, indicating their disproportionately high impact on the information ecosystem. Modern bots utilize adversarial machine learning techniques to evade detection. They simulate variable activity patterns, utilize diverse linguistic styles, integrate authentic content with disinformation, and create networks of mutually reinforcing accounts, making identification significantly more difficult.

AI algorithms enable microtargeting, delivering personalized disinformation content to precisely defined user segments. Analysis of demographic, psychographic, and behavioral data allows for the creation of messages that maximize virality within specific target groups. These techniques, originally used in commercial marketing, have been adapted for disinformation campaigns. Of particular concern is the rise of "dynamic disinformation"—content generated in real time in response to user interactions, location, and the current social and political context. These systems use generative models to create unique variations of underlying disinformation narratives, optimized to maximize persuasion for a specific audience.

Social media platforms' algorithms, designed to maximize engagement, naturally promote controversial and emotional content, regardless of its accuracy. This mechanism creates a positive feedback loop: disinformation generates high engagement, and algorithms increase its reach, leading to further engagement. In the post-truth era, AI algorithms unintentionally create echo chambers that amplify disinformation. [16]Users are primarily exposed to content that aligns with their existing beliefs, leading to polarization and reduced resistance to disinformation. Research shows that exposure to diverse perspectives decreases by approximately 25-35% in algorithmically curated feeds compared to chronologically ordered content.

Ethical, social and technological risks

One of the fundamental challenges in automatic disinformation detection is the problem of false positives. Errors in predictive models lead to content being falsely flagged as disinformation, restricting freedom of speech. Satire, contextual opinions, rhetorical exaggerations, and ironic content can be misclassified by AI systems lacking a full understanding of cultural context and communicative intent. Research by Mouratidis and colleagues (2025) has shown that even advanced models based on Transformer architectures achieve a false positive rate (FPR) of 8-15%, depending on the content category [13]. In practice, this means that with millions of posts analyzed daily, hundreds of thousands of legitimate content items are incorrectly flagged as potential disinformation. This problem is particularly significant in the context of political content, where the line between legitimate criticism and disinformation is often ambiguous.

Most disinformation detection systems are developed and trained on English data, leading to significant performance differences across languages. NLP models exhibit a 30-50% accuracy drop when applied to low-resource languages with limited training sets. This problem has key global implications: disinformation in non-English languages, including Slavic, Asian, and African languages, is significantly more difficult to automatically detect. Furthermore, cultural differences in rhetoric, humor, and forms of communication mean that models trained on data from one cultural context exhibit limited transferability to other contexts.

While text-based disinformation detection has reached a relatively high technological maturity, detecting multimodal disinformation (combining text, images, video, and audio) remains a significant challenge. Audio and video deepfakes require specialized detection techniques that are often vulnerable to adversarial attacks - intentional content modifications intended to deceive detection systems. Current deepfake detection systems based on analysis of compression artifacts, lighting inconsistencies, and eye movement anomalies achieve accuracy of around 85-90% in controlled laboratory conditions, but their accuracy drops to 60-70% in real-world conditions, where varying material quality, different social media platform compression techniques, and intentional obfuscation techniques arise.

AI systems are susceptible to deliberate manipulations aimed at bypassing detection mechanisms. Adversarial attacks in the context of disinformation detection include subtle content modifications (e.g., replacing individual letters with visually similar Unicode characters, introducing noise into images, or stylistic modifications to text) that are invisible to humans but cause models to misclassify. Prompt injection, in the case of systems using LLMs for content moderation, involves inserting instructions into the analyzed text intended to manipulate the classification process. Research shows that approximately 40-60% of advanced LLM-based systems are vulnerable to this type of attack, posing a significant threat to their credibility as moderation tools.

Most advanced disinformation detection models rely on deep neural networks, which function as "black boxes", their decisions are difficult to interpret, even for experts. The lack of transparency in the classification process makes it difficult to verify the accuracy of decisions, identify sources of error, and build user trust in the systems. This issue is particularly relevant in the context of content moderation, where users have the right to understand why their content was flagged as disinformation. The development of Explainable AI (XAI) techniques for disinformation detection is currently an active area of research, but existing solutions (e.g., LIME, SHAP) provide only approximate, often incomplete explanations of model decisions.

Processing billions of posts, comments, images, and videos generated daily on social media platforms in near real time presents a significant infrastructure challenge. Advanced deep learning models require significant computational resources, limiting their applicability across entire platforms. The trade-off between accuracy and processing speed is a fundamental challenge: simpler, faster models achieve lower accuracy, while more advanced, higher-accuracy models are too slow for real-time use. Social media platforms often employ a multi-stage approach, where fast, less accurate models perform initial screening, while more advanced models analyze only content flagged as suspicious in the first stage.

F. Real-world implementations and case studies of effective interventions

The COVID-19 pandemic has posed an unprecedented test for misinformation detection systems, generating an "infodemic" , a surge in both medical information and misinformation. According to an analysis by Renda and Simonelli (2025), major social media platforms have integrated AI with fact-checking systems, which has reduced users' exposure to medical misinformation [22]. Facebook/Meta implemented a system combining automatic detection based on BERT-based models with verification by independent fact-checkers. The system achieved the following results: (1) approximately 180 million posts were flagged as containing potentially false information about COVID-19; (2) future views of flagged content were reduced by an average of 95%; (3) information overlays were added to 50 million posts containing partially true but potentially misleading information. YouTube implemented algorithms that promote authorized medical sources (WHO, CDC, national public health authorities) in search results and recommendations. The analysis found that videos from authorized medical channels saw a 10-fold increase in views during the pandemic compared to the pre-pandemic period, while content containing medical misinformation experienced a drop in reach of around 70%.

Experimental systems integrating AI with blockchain technology are developing the concept of "trusted fact databases." Blockchain provides an immutable record of verified facts, while AI models automate the process of verifying new claims against this database. [15] The Duke Reporters' Lab project developed the ClaimReview schema, a standardized metadata format for verified facts that has been adopted by major search engines. Integrating this standard with AI algorithms allows for automatic cross-referencing of new claims with a database of verified facts, significantly speeding up the fact-checking process. An experimental implementation of the FactChain blockchain system demonstrated that the immutable nature of distributed ledger technology can increase trust in fact-checking results by approximately 40% among users skeptical of centralized moderation platforms. However, the system faces scalability challenges: blockchain consensus verification introduces latency of several minutes, which is unacceptable for real-time detection.

The most promising results in combating disinformation are achieved by hybrid systems that combine automated AI detection with expert human verification. The VERA.ai (Verification of Real-time Anonymized Information) project exemplifies this approach, integrating NLP algorithms with crowdsourcing of independent experts and investigative journalists [21]. The system operates in three stages: (1) AI algorithms perform an initial screening, identifying potentially problematic content based on linguistic patterns and online behavior; (2) content flagged as suspicious is referred for verification by experts; (3) verified facts feed the training database for AI models, creating a continuous learning loop. This architecture achieves an accuracy of around 94% while maintaining scalability - the system processes millions of posts per day, while experts only verify about 0.5% of the content flagged as most problematic. A study by Pennycook and Rand (2019) found that crowdsourced judgments of ordinary users regarding the quality of news sources correlate highly (r=0.82) with those of professional fact-checkers, suggesting potential for building systems based on the "wisdom of the crowd" [21]. Algorithms that aggregate multiple user judgments can achieve accuracy comparable to that of individual experts, while reducing costs and increasing scalability.

G. Regulatory Framework and Public Policy Recommendations

The European Union has adopted the most comprehensive regulatory approach to disinformation and the role of AI in its moderation. The Digital Services Act (DSA), which entered into force in 2024, requires very large online platforms (VLOPs) to assess the systemic risk associated with the spread of disinformation and implement proportionate mitigation measures. The DSA requires transparency in recommendation algorithms and allows users to access versions of platforms that do not rely on algorithmic personalization. Early assessments of the DSA's effectiveness are mixed: platforms report difficulties in meeting transparency requirements while maintaining commercial confidentiality, while regulators point to insufficient mitigation measures implemented by the platforms. The AI Act, also originating from the EU, classifies AI systems used in content moderation as “high risk” requiring rigorous testing, documentation, and human oversight. This regulation potentially raises the quality standards of disinformation detection systems, but it could also increase entry barriers for smaller platforms and innovative solutions.

In the United States, there is no comprehensive federal regulation governing content moderation by social media platforms. Section 230 of the Communications Decency Act grants platforms broadly legal immunity for user-generated content, which critics argue reduces their incentive to aggressively combat disinformation. At the same time, Section 230 shields platforms from liability for their moderation decisions, paradoxically allowing them to take action against disinformation without risking lawsuits from users. Regulating AI in the context of disinformation detection faces fundamental challenges stemming from the nature of technology. The "black box" problem makes it difficult for regulators to assess whether AI systems are operating in accordance with their stated principles and are not engaging in undue censorship. Algorithmic transparency requirements are difficult to implement without disclosing implementation details, which constitute the platforms' commercial value and could also be exploited by disinformation actors to bypass detection systems. The rapid pace of AI technology development means that regulations quickly become obsolete. Next-generation language models emerge in a cycle of several months, radically transforming both the capabilities of disinformation generation and detection. Traditional legislative processes, which take years, cannot keep up with this pace of change. The issue of jurisdiction poses an additional challenge in the global information ecosystem. Disinformation often transcends national borders, while regulations are typically national or regional. Social media platforms operate globally but must adapt to varying regulatory requirements across jurisdictions, leading to fragmentation of their moderation systems and potentially unequal treatment of users across countries. [26]

Based on the analysis of literature and case studies, it is possible to formulate a set of recommendations for social media platforms striving to effectively combat disinformation while respecting freedom of speech and user privacy. Effective public policy to combat disinformation requires a balanced approach that takes into account diverse values and interests:

- Risk-based regulatory frameworks: (1) Classify AI systems by risk level with proportionate regulatory requirements - the highest standards for systems used in policy and public health contexts. (2) Require impact assessments before implementing AI systems with high societal impact, analogous to environmental impact assessments. (3) Establish independent oversight bodies with technical competence to evaluate AI systems and enforce standards.
- Investments in education and media literacy: (1) Media and digital literacy programs in school systems that teach critical evaluation of information sources and recognition of manipulation techniques. (2) Public awareness campaigns about deep-fakes, social bots, and other disinformation techniques. (3) Support for investigative journalism and fact-checking through public grants and tax breaks.
- Support for research and innovation: (1) Funding academic research on disinformation detection, with a focus on less resourced languages and cultural contexts underrepresented in current systems. (2) Creating publicly available training datasets for the development of detection tools, while maintaining privacy standards. (3) Public-private initiatives combining technology platform resources, academic expertise, and regulatory mandates.
- Protecting freedom of speech and pluralism: (1) Safeguards against the use of anti-disinformation systems to censor legitimate criticism and political opposition. (2) Appeal mechanisms and due process for users whose content has been modified. (3) Special precautions in electoral contexts, where the line between disinformation and political polemics is most ambiguous. [27]

NGOs, think tanks, and research groups play a key role in the anti-disinformation ecosystem, acting as a bridge between technology platforms, regulators, and society. The European Digital Media Observatory (EDMO) is an example of an effective network that brings together fact-checkers, researchers, and regulators to coordinate counter-disinformation efforts across the EU. [22] Independent monitoring and watchdog functions by civil society organizations are essential to ensuring the accountability of platforms and regulators. Projects like Algorithm Watch document instances of mismanagement and the discriminatory effects of algorithms, putting pressure on platforms to improve their systems. Crowd-sourced fact-checking initiatives, such as Wikipedia's approach and Community Notes on X, demonstrate the potential of bottom-up, community-based fact-checking mechanisms. Research shows that such approaches can achieve high accuracy while maintaining greater social acceptance than top-down moderation imposed by platforms or governments.

H. Development prospects and future challenges

The fight against disinformation in the AI era is characterized by an "arm race" dynamic; each advance in detection technologies is matched by the development of more advanced techniques for generating and obfuscating disinformation. The emergence of GPT-4 and Claude models significantly raised the bar for generating convincing text-based disinformation, while diffusion models (Stable Diffusion, DALL-E 3, Midjourney) made the creation of false images accessible to everyone. [28] Upcoming multimodal models (GPT-4V, Gemini Ultra), combining text, image, audio, and video understanding, will likely enable the generation of complex, multidimensional disinformation campaigns with unprecedented coherence and persuasion. Detecting such campaigns will require similarly advanced multimodal systems, which will increase the computational and financial requirements for effective moderation.

Artificial intelligence in the context of disinformation is a classic example of dual-use technology, technology that can serve both beneficial and harmful purposes [1]. This fundamental property complicates regulatory strategies and requires a nuanced approach that accounts for trade-offs. Completely halting the development of generative AI to prevent its misuse would be unrealistic and undesirable, given the numerous beneficial applications of these technologies in education, creativity, scientific productivity, and other areas. At the same time, unrestricted development and democratization of access to the most advanced generative models could significantly lower the barriers to entry for disinformation actors. Some experts advocate a "responsible disclosure" model for advanced AI systems, analogous to cybersecurity practices, where details of critical vulnerabilities are disclosed with a delay allowing for the development of countermeasures. Others argue that attempts to control the distribution of AI technologies are doomed to failure in the face of open implementations and international competition.

In the long term, the proliferation of deepfakes and AI-generated disinformation could lead to a fundamental transformation of societal trust and epistemology. The concept of "infocalypse" or "epistemic crisis" describes a scenario in which the inability to distinguish authentic content from synthetic content leads to an erosion of trust in all digital information. Paradoxically, this situation could also stimulate the development of new verification and authentication mechanisms. Technologies such as cryptographic signatures for digital media (e.g., the Coalition for Content Provenance and Authenticity (C2PA) standard), watermarking for AI-generated content, and blockchain-based proof-of-concept tracking could evolve into standard trust infrastructures for digital content. Some researchers suggest that societies could develop increased "epistemic resilience" through adaptive learning in a disinformation-saturated environment, analogous to the development of immune resilience. However, such adaptation would require significant investment in media literacy and critical thinking and could occur unevenly across demographic groups, potentially exacerbating existing social divisions.

Integration of new technologies: quantum computing and neurotechnology. Future advances in quantum computing could radically shift the balance of power in the fight against disinformation. Quantum computers could potentially break current cryptographic systems that protect the authenticity of digital content, while also offering the potential for developing new, more advanced methods for detecting patterns in massive datasets. Emerging neurotechnologies, such as brain-computer interfaces, open up entirely new frontiers for disinformation and manipulation. A direct interface between AI systems and the human brain could enable forms of influence and manipulation that go beyond the current capabilities of social media, requiring fundamentally new ethical and regulatory frameworks.

I. Conclusions

An analysis of the role of AI in predicting content virality and combating disinformation reveals the deeply ambivalent nature of these technologies [13]. On the one hand, AI is a powerful tool enabling scalable detection and mitigation of disinformation - predictive models achieve 78-92% accuracy in identifying potentially viral false content [29], [19], hybrid systems combining automation with expert verification demonstrate effectiveness in real-world implementations, and the integration of AI with fact-checking has contributed to a measurable reduction in exposure to disinformation in critical contexts such as the COVID-19 pandemic [22]. On the other hand, these same technologies are fundamentally changing the disinformation landscape, lowering the barriers to entry for producing convincing false content through generative AI and deepfakes, [23] enabling mass automation and scaling of disinformation campaigns through social bots, [24] and creating new vectors of manipulation through microtargeting and personalization. This duality characterizes AI as the quintessential dual-use technology in the information context.

Current AI-based disinformation detection systems face significant technical and conceptual limitations. The problem of false positives (8-15% even for advanced models) threatens freedom of speech and can lead to excessive censorship of legitimate content. [13] Scalability challenges limit the ability to conduct in-depth verification to a fraction of a percent of the total volume of content generated on social media platforms. The limited interpretability of deep learning models hinders accountability and building trust in moderation systems. Disparities in effectiveness across languages and cultural contexts lead to unequal protection against disinformation, with communities speaking less resourced languages and regions with limited representation in training sets particularly vulnerable. The vulnerability of AI systems to adversarial attacks and prompt injections ensures that the arms race between disinformation detection and generation will continue.

Effectively combating disinformation in the AI era requires a multifaceted approach that goes beyond purely technological solutions. Integrating AI technologies with human expert verification in hybrid systems combines the scalability of automation with the accuracy and contextual understanding of human judgment. [21] Collaboration between technology platforms, regulators, academia, and civil society is essential for coordinating efforts and exchanging knowledge. Media education and civic literacy constitute a fundamental defense against disinformation, building societal resilience independent of the effectiveness of technical systems. A balanced regulatory framework, based on risk and proportionality, can stimulate the development of responsible technologies while protecting fundamental democratic values. Investment in research and development, particularly in the areas of multilingualism, multimodal sensing, and explainable AI, is crucial for long-term effectiveness.

This analysis identifies several areas requiring further research. First, longitudinal studies are needed to assess the long-term effectiveness of disinformation detection systems and their impact on information ecosystems. Second, comparative studies across jurisdictions and regulatory frameworks can reveal best practices and lessons learned. Third, developing explainable AI techniques specifically for disinformation detection contexts is crucial for accountability and public trust. Research on social perception and trust in AI systems, the psychology of disinformation in the era of deepfakes, and mechanisms for building social resilience to information manipulation are complementary areas requiring an interdisciplinary approach combining computer science, psychology, sociology, and political science. Finally, proactive research on emerging technologies such as quantum computing and neurotechnology in the context of disinformation can enable anticipatory governance, rather than reactively responding to already materialized threats. The way forward requires balancing technological optimism with a realistic recognition of limitations, combining innovation with responsibility, and prioritizing human flourishing and democratic values over purely technical performance metrics. Artificial intelligence can be a powerful ally in protecting information integrity, but only if developed and implemented with wisdom, ethics, and democratic oversight [1].

Conclusion

AI has moved beyond being a simple classroom tool. It is now an infrastructural force that dictates how knowledge is shared and valued. This article has tried to show that what happens in a lecture hall, like AI-generated assignments, cannot be separated from what happens in the wider information ecosystem. While these systems can make learning feel more personal in the short term, they also carry a hidden cost: a growing cognitive dependency that can quietly hollow out critical thinking.

Our analysis suggests that the real risk lies in how AI-driven algorithms prioritize 'viral' content over actual epistemic quality. By linking educational technology with information studies, we’ve highlighted a feedback loop of automation that traditional research often misses. However, we must acknowledge that this paper is built on a conceptual synthesis rather than large-scale empirical data. Given the "moving target" nature of AI development, our ability to generalize these findings is necessarily limited. To truly understand the long-term impact, future work must move beyond theory and toward longitudinal studies that can track student development over years, not just months. We also need more data-driven analyses of how information actually diffuses through these automated ecosystems. The takeaway is that we cannot just 'plug in' AI and hope for the best. Strengthening academic literacy and critical engagement remains our only real defense against the risks of automation.

References

[1] M. W. Romaniuk, J. Gierzyński, M. M. Pietrzak and J. Zbróg, "Integrating Technology in Social Science Research: Emerging Trends and Ethical Considerations," International Journal of Electronics and Telecommunications, vol. 71, no. 1, pp. 171-179, 2025.

[2] M. W. Romaniuk, A. Szarfenberg, I. Pawłowska and K. Choszczyk, "Doctoral Theses in the Digital Age – ICT use by Social Sciences PhD Students of The Maria Grzegorzewska University," International Journal of Electronics and Telecommunications, vol. 70, no. 1, pp. 199-204, 2024.

[3] M. W. Romaniuk, P. Mika, J. Apanasewicz and E. Duda-Maciejewska, "Enhancing Research Practices: Digital Technologies in the Social Sciences and Practical Tools for Doctoral Students," International Journal of Electronics and Telecommunications, vol. 71, no. 1, pp. 181-188, 2025.

[4] S. a. S. Z. M. O. Zakaria, "STEAM innovation: Curriculum alignment, experimental learning, and transdisciplinary approaches.," International Journal of Modern Education 6.22, pp. 319-335, 2024.

[5] J. M. Y. J. M. H. A. A. S. U. &. H. S. Jawaid, "Robotic system education for young children by collaborative-project-based learning.," Computer Applications in Engineering Education, 28(1), pp. 178-192, 2020.

[6] M. Á. R. F. J. F. C. G. J. L. J. &. G. F. J. Conde, "Fostering STEAM through challenge-based learning, robotics, and physical devices: A systematic mapping literature review.," Computer Applications in Engineering Education, 29(1), pp. 46-65, 2021.

[7] Y.-F. Y. Y.-W. C. N.-S. C. Chih-Chien Hu, "Integrating educational robot and low-cost self-made toys to enhance STEM learning performance for primary school students," Behavior & Information Technology, pp. 1614-1635, 2024.

[8] P. C. (2022), "Project-Based STEM Learning Using Educational Robotics as the Development of Student Problem-Solving Competence.," Mathematics, p. 4618, 2022.

[9] Y. J. &. L. K. Y. Sie, "Effects of Psychological Capital and Cognition on STEM Learning in IoT Smart Energy-Saving Project," Journal of Baltic Science Education, 24(2), pp. 340-359, 2025.

[10] C. N. P. C. C. &. T. D. Ferreira, "Socio-constructivist teaching powered by ICT in the STEM areas for primary school.," 12th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-5, 2017.

[11] S. Y. L. C. C. &. S. J. Y. Lu, "Project-based learning oriented STEAM: The case of micro–bit paper-cutting lamp.," International Journal of Technology and Design Education, 32(5), pp. 2553-2575, 2022.

[12] V. &. P. F. Karampa, "A motivational design of a flipped classroom on collaborative programming and STEAM. In," International Workshop on Learning Technology for Education in Cloud. Cham: Springer International Publishing, pp. 226-238, 2018.

[13] D. Mouratidis, A. Kanavos, K. Kermanidis, "From Misinformation to Insight: Machine Learning Strategies for Fake News Detection," Information, vol. 16, p. 189, 2025.

[14] Sajjad, M., Kwon, Y. I., Lee, S., "A novel CNN-GRU approach for user independent emotion recognition from speech," Information Processing & Management, vol. 57, 2020.

[15] Krstovski, K., Ryu, A. S., & Kogut, B., "Evons: A dataset for fake and real news virality analysis and prediction," arXiv, 2022.

[16] C. &. D. H. Wardle, "Information Disorder: Toward an interdisciplinary framework for research and policymaking," 2017.

[17] Berger, J., & Milkman, K. L., "What makes online content viral?," Journal of Marketing Research, vol. 49, pp. 192-205, 2012.

[18] Vosoughi, S., Roy, D., & Aral, S., "The spread of true and false news online," Science, vol. 359, pp. 1146-1151, 2018.

[19] Gondwe, G., "Can AI outsmart fake news? Detecting misinformation with AI models in real-time," Emerging Media, 2025.

[20] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H., "Fake news detection on social media: A data mining perspective," Journal of Big Data, vol. 4, p. 38, 2017.

[21] Pennycook, G., & Rand, D. G., "Fighting misinformation on social media using crowdsourced judgments of news source quality," Proceedings of the National Academy of Sciences, vol. 116, pp. 2521-2526, 2019.

[22] Dierickx, L., Lindén, C.-G., & Dang-Nguyen, D.-T., "Part of the problem and part of the solution: the paradox of AI in fact-checking," 2025.

[23] Sedova, K., McNeill, C., Johnson, A., Joshi, A., & Wulkan, I., "AI and the Future of Disinformation Campaigns: Part 2 – A Threat Model," Center for Security and Emerging Technology, 2021.

[24] Giroux, J., Ariyarathne, G., Nwala, A. C., & Fanelli, C., "Unmasking social bots: How confident are we?," EPJ Data Science, vol. 14, p. 18, 2025.

[25] Klepper, D., & Swenson, A., "AI-generated disinformation poses threat of misleading voters in 2024 election," PBS NewsHour, 2024.

[26] Posłajko, R., "Rozwój europejskiej polityki cyfrowej," Poliarchia, vol. 2, pp. 37-61, 2014.

[27] Gomez, J. F., Machado, C., Paes, L. M., & Calmon, F., "Algorithmic Arbitrariness in Content Moderation," Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024.

[28] López-Borrull, A., & Lopezosa, C., "Mapping the impact of generative AI on disinformation: Insights from a scoping review," Publications, vol. 13, p. 33, 2025.

[29] Wang, H., Chen, H., & Zhang, Y., "An end-to-end joint model for evidence information extraction from court record document," Information Processing & Management, vol. 57, 2020.