Journal of computer science and software applications

https://www.mfacademia.org/index.php/jcssa

ISSN:2377-0430
Vol. 5, No. 6, 2025

A Survey on AI-Generated Content: From Technical Foundations to Ethical Challenges

Callum Redgrave
Lakehead University, Thunder Bay, Canada
callum.r875@lakeheadu.ca

Abstract:

The emergence of AI-generated content (AIGC) technologies has transformed the landscape of digital media, communication, and knowledge generation. Driven by the rapid evolution of foundation models such as GPT, DALL·E, and Stable Diffusion, AIGC systems have demonstrated remarkable capabilities in generating human-like text, images, code, and multimedia. This paper provides a comprehensive review of the technical underpinnings, system-level architectures, and deployment workflows that support modern AIGC systems. It further explores critical ethical and societal concerns, including misinformation, authorship attribution, bias propagation, and the erosion of human agency. Through an examination of recent developments from 2023 to 2025, the paper evaluates the trajectory of AIGC integration across sectors such as education, entertainment, law, and journalism. Finally, we discuss emerging regulatory frameworks and outline key research directions to ensure responsible, transparent, and human-aligned advancement of generative AI.

Keywords:

AI-generated content (AIGC), generative models, large language models (LLMs), diffusion models, reinforcement learning with human feedback (RLHF), content moderation

1. Introduction

Artificial Intelligence Generated Content (AIGC) represents a significant technological milestone in the evolution of computational creativity and human-computer interaction. Built upon advances in deep learning, natural language processing, and generative modeling, AIGC systems now possess the ability to produce text, images, videos, code, and even music with a degree of fluency and semantic coherence previously unattainable by machines. The widespread adoption of transformer-based architectures—particularly large-scale language and diffusion models—has catalyzed a paradigm shift in how content is created, distributed, and consumed.

The commercial success of tools such as OpenAI's ChatGPT, Midjourney, and Google's Gemini has accelerated the integration of AIGC across industries. From automated content generation in digital marketing to virtual teaching assistants in education, these systems have shown potential to enhance productivity, personalize experiences, and democratize creative expression. At the same time, they pose serious ethical and societal risks, including the spread of disinformation, deepfake creation, algorithmic bias, and challenges to traditional notions of intellectual property and authorship [1][2].

The technical foundation of AIGC lies in a combination of pretraining on massive corpora, fine-tuning for domain-specific tasks, and deployment in latency-sensitive environments. Architectures such as encoder-decoder transformers, latent diffusion models, and reinforcement learning with human feedback (RLHF) form the backbone of current generative systems. Alongside these models, a growing ecosystem of prompt engineering strategies, content filters, inference APIs, and open-source toolkits enables the scalable delivery of AIGC services.

Despite the rapid progress, the deployment of AIGC remains fraught with open challenges. Technical issues such as hallucination, adversarial prompt exploitation, model collapse under continual fine-tuning, and the compute costs of large model inference require ongoing research. Ethically, the lack of transparency in model decisions, the difficulty in detecting synthetic content, and the disproportionate influence of training data on generative outcomes raise questions about accountability and fairness. Moreover, geopolitical concerns about surveillance, manipulation, and digital sovereignty have led to calls for regulatory intervention and the development of safety-aligned governance frameworks [3][4].

This survey aims to provide a comprehensive and structured overview of the AIGC domain from both technical and ethical perspectives. It is organized as follows: Section 2 presents the evolution and core architecture of generative models powering AIGC. Section 3 analyzes the system pipeline of AIGC deployment, including data pipelines, model optimization, and content serving. Section 4 reviews application scenarios across industries. Section 5 focuses on ethical challenges, including model misuse, fairness, transparency, and value alignment. Section 6 outlines emerging mitigation strategies and policy frameworks. Finally, Section 7 highlights key research directions that could shape the responsible future of AIGC.

2. Generative Model Architectures for AIGC

The foundation of modern AI-generated content (AIGC) lies in large-scale generative models that are capable of learning complex data distributions and generating high-quality outputs across various modalities. These models have evolved through a combination of architectural innovation, training scale, and optimization strategies. The dominant paradigms include transformer-based language models, diffusion-based generative models, and reinforcement learning with human feedback (RLHF), each contributing uniquely to the capabilities and behaviors of AIGC systems.

Transformer-based architectures have emerged as the most impactful innovation in natural language generation. Originally introduced by Vaswani et al. in 2017, the transformer model discards recurrence and convolutions in favor of a self-attention mechanism that allows for efficient parallelization and long-range dependency modeling. Subsequent iterations such as BERT, GPT, T5, and LLaMA expanded these concepts by scaling model size and leveraging masked or autoregressive objectives for pretraining. GPT-style models, in particular, have demonstrated exceptional performance in zero-shot and few-shot tasks due to their autoregressive token generation capability and massive pretraining corpora [5]. By 2025, models like GPT-4, Gemini 1.5, Claude 3, and open-source competitors such as Mistral and Mixtral have achieved human-comparable fluency in long-form text generation, reasoning, and code synthesis [6].

For image, audio, and video generation, diffusion-based generative models have gained prominence. These models operate by learning to reverse a gradual noise-injection process, effectively denoising a latent representation into a coherent data sample. Stable Diffusion, DALLE-2, and Imagen are representative examples that use text-conditioned diffusion processes to generate high-fidelity visuals with controllable attributes. Compared to earlier generative adversarial networks (GANs), diffusion models offer more stable training dynamics and greater diversity in output. The flexibility of conditioning mechanisms—whether via text, sketches, or semantic maps—has enabled diffusion models to support multi-modal content generation, image inpainting, and style transfer tasks with unprecedented visual realism [7][8].

In recent years, reinforcement learning with human feedback (RLHF) has become an essential technique for aligning model behavior with human expectations. While pretraining provides models with linguistic competence, it does not inherently align outputs with values such as truthfulness, helpfulness, or harmlessness. RLHF addresses this by incorporating human preferences into the reward function of the model, using comparative judgments to train a reward model and applying policy optimization to fine-tune the base model. This approach has been critical to the safety and usability of models like ChatGPT and Claude, which rely on human-in-the-loop learning to discourage toxic, biased, or nonsensical outputs [9]. Research has also extended RLHF to multi-objective optimization, where alignment is balanced with creativity or domain-specific knowledge.

In parallel, retrieval-augmented generation (RAG) techniques have been developed to address the limitations of static model memory. By augmenting generative models with access to external knowledge bases or document stores, RAG systems can dynamically retrieve relevant context and condition their outputs on it. This not only reduces hallucination but also enables models to stay updated without retraining. Techniques such as kNN-LM, RETRO, and vector database integration are widely used in enterprise AIGC deployments where factual consistency and traceability are critical [10].

Across these architectures, model scaling remains a key trend. Larger models generally exhibit emergent abilities such as translation, summarization, and commonsense reasoning. However, increased size also introduces challenges in training cost, inference latency, and carbon footprint. To address this, recent research has explored efficient model variants, such as sparsely activated Mixture-of-Experts (MoE), quantized transformers, and low-rank adaptation (LoRA), which enable parameter-efficient fine-tuning and deployment at scale. Moreover, modular architectures that separate capability into specialized expert modules offer a promising direction for flexible and controllable generation [11].

In summary, the technical architecture of AIGC systems reflects a convergence of innovations in sequence modeling, diffusion processes, reinforcement learning, and external retrieval integration. These components jointly empower modern generative models to produce coherent, high-quality, and context-sensitive content across modalities. Understanding their design principles is crucial not only for improving performance but also for diagnosing risks, developing alignment strategies, and building transparent, trustworthy AIGC systems.

3. System Pipeline and Deployment Architecture of AIGC

The deployment of AI-generated content (AIGC) systems involves more than just training large generative models; it requires a comprehensive system pipeline that spans data acquisition, model optimization, infrastructure provisioning, inference acceleration, and user interaction. Unlike traditional machine learning workflows, AIGC pipelines are typically designed for low-latency, high-throughput, and multimodal generation tasks, often deployed in heterogeneous environments including cloud services, edge devices, and integrated APIs. This section outlines the end-to-end architecture of AIGC systems, identifying key technical components and engineering trade-offs that influence performance, scalability, and user safety.

At the foundation of any AIGC system is the data preprocessing and model training stage. For large language and diffusion models, pretraining requires ingesting massive, diverse, and often noisy datasets, including web crawls, books, academic papers, code repositories, and social media content. This phase includes text normalization, de-duplication, filtering of unsafe content, and data balancing to mitigate the overrepresentation of specific domains or linguistic biases. The quality and diversity of training data are directly correlated with the generative fluency and generalization capability of the resulting model. Recent best practices involve combining proprietary curated corpora with open-domain internet data, followed by staged pretraining and supervised fine-tuning [12].

Following model training, the pipeline progresses into inference optimization and deployment engineering. Modern AIGC models contain billions of parameters and are typically stored as large checkpoint files that require significant memory bandwidth and GPU compute to serve in real time. To reduce inference latency and hardware costs, techniques such as quantization, model distillation, parameter-efficient fine-tuning (e.g., LoRA), and early exit strategies are commonly applied. Furthermore, serving frameworks such as DeepSpeed, FasterTransformer, and vLLM have been developed to enable high-throughput, batched, and streaming inference across GPU clusters [13]. The choice of batch size, precision (FP16, INT8), and caching strategies has direct implications for model responsiveness in user-facing applications.

A critical element in the AIGC architecture is the serving stack and interface layer, which defines how users interact with generative models. Most systems expose functionality through RESTful APIs or socket-based streaming protocols. These interfaces typically sit behind orchestration layers that manage user authentication, rate limiting, session management, and content moderation. Companies such as OpenAI, Anthropic, and Google have integrated middleware components that apply filters for unsafe outputs, enforce usage policies, and log interactions for post-hoc auditing. Contextual prompt pre-processing and post-generation safety checks form part of a real-time feedback loop designed to minimize harmful outputs and ensure responsible usage [14].

In multi-turn conversational systems and multi-modal pipelines, session memory and tool invocation capabilities add architectural complexity. Systems like ChatGPT and Gemini leverage external memory modules or vector stores to maintain long-term conversational coherence. Some also integrate plugins or agents that allow the model to call external APIs (e.g., calculators, web search, databases) during generation. These tool-use extensions are orchestrated through planning modules and response parsers that convert high-level user intent into structured calls and natural language outputs. As such, the pipeline becomes a composite of deterministic logic and probabilistic generative modules, requiring careful coordination and failure recovery strategies [15].

Scalability and robustness are major concerns in production environments. Enterprises often rely on Kubernetes-based deployment stacks, containerized runtime environments, and cloud auto-scaling to handle traffic spikes. Observability tools are used to monitor token usage, request latency, rejection rates, and hardware health. To support continuous learning and personalization, some systems implement user feedback loops and reinforcement-based fine-tuning pipelines. However, such updates must be carefully managed to avoid model collapse, drift, or reinforcement of biased behaviors.

Finally, deployment must consider geographic compliance, data sovereignty, and privacy policies. Depending on regulatory jurisdictions (e.g., GDPR in Europe, CCPA in California, CAC rules in China), AIGC systems may need to localize inference endpoints, redact personally identifiable information (PII), or disable logging for sensitive content. Data anonymization and auditability are thus not optional but integral to system architecture.

In summary, AIGC system pipelines integrate diverse components across the ML lifecycle, cloud infrastructure, inference engineering, and safety governance. The architecture must balance competing requirements: speed vs. safety, generality vs. specialization, and scale vs. personalization. As deployment moves from research labs to consumer-grade applications, the engineering discipline behind AIGC systems becomes central to ensuring reliability, accountability, and user trust.

4. Applications of AIGC Across Domains

AI-generated content (AIGC) technologies have rapidly permeated a wide array of sectors, extending their impact well beyond the confines of research laboratories into real-world production systems. With capabilities to generate natural language, visual media, code, and synthetic speech, AIGC tools are increasingly used to augment or automate tasks traditionally performed by human professionals. The cross-domain adoption of AIGC is being driven by its potential to increase efficiency, reduce costs, personalize content, and scale creativity. However, the transformative potential of these tools also brings disruption, particularly in domains that rely on authenticity, accuracy, or human judgment.

In journalism and media, AIGC is being employed to automate the generation of routine articles such as weather reports, financial updates, and sports summaries. News agencies like the Associated Press and Bloomberg have adopted template-driven AIGC systems that generate thousands of articles per day using structured data feeds. More advanced systems powered by large language models (LLMs) have been used to assist in headline generation, content summarization, and even exploratory writing for investigative journalism. However, the use of AIGC in newsrooms has raised questions about factual reliability, source attribution, and the risk of “hallucinated” information being published as truth. A growing trend involves integrating retrieval-augmented generation (RAG) into editorial workflows to ground model outputs in verifiable documents [16].

In education, AIGC has been adopted for personalized tutoring, content generation, and automated feedback. Tools like ChatGPT, Gemini, and Socratic are used to generate explanations, quiz questions, code snippets, and writing prompts tailored to the learner’s profile. These models can adapt their responses to different levels of student proficiency, offering scalable and on-demand assistance. Universities are also using AIGC to support curriculum development, lecture transcript generation, and virtual teaching assistants. While promising, concerns remain about overreliance on machine-generated explanations, potential inaccuracies in scientific reasoning, and the erosion of students’ critical thinking skills. Educators must design usage frameworks that blend AIGC with pedagogical oversight [17].

In the legal domain, AIGC systems are being used for contract drafting, legal summarization, and statutory interpretation. Legal technology companies have integrated LLMs into document review workflows to identify clauses, summarize case law, and suggest revisions. Some applications even assist in legal research by generating structured summaries from court rulings. However, the interpretability and accuracy of these systems remain contested. High-stakes decisions in litigation, compliance, and due diligence require human validation, as errors in legal reasoning could have serious consequences. Bar associations and regulators have begun to issue guidelines around the responsible use of generative AI in legal practice [18].

In creative industries, AIGC has enabled new forms of artistic expression. Visual artists use diffusion-based models such as Midjourney and DALL·E to create concept art, storyboards, and design prototypes. Writers and filmmakers leverage LLMs for dialogue generation, narrative ideation, and script drafting. In the music industry, AI systems can now compose melodies, simulate vocals, and remix tracks in real time. This democratization of creativity has empowered individuals without formal training to produce professional-grade content. Yet it also raises contentious issues around originality, authorship, and the displacement of human creators. Intellectual property laws and licensing frameworks are being tested as courts consider whether AI-generated works can be copyrighted or attributed to human collaborators [19].

In software development, AIGC systems such as GitHub Copilot, Amazon CodeWhisperer, and Replit AI have revolutionized coding practices by offering real-time code suggestions, bug fixes, and even full-function generation. Developers benefit from increased productivity and reduced context-switching during coding. In enterprise settings, these tools assist in legacy code migration, documentation, and compliance automation. However, risks include the unintentional reuse of copyrighted code from training data, the propagation of insecure coding patterns, and developer overdependence on auto-generated logic. To mitigate these concerns, research on code attribution, static analysis of AI-generated code, and human-in-the-loop coding environments is gaining traction [20].

AIGC also plays an emerging role in scientific research and publishing. Language models have been applied to generate literature reviews, summarize experimental results, and even assist in hypothesis formation. Some preprint servers and journals now use AI systems to help screen submissions for quality and relevance. Scientific visualization tools driven by generative models support automated figure creation and data narration. Nonetheless, the possibility of fabricating results or plagiarizing prior work using AIGC has raised ethical alarms, leading to calls for transparent disclosure of AI assistance in scholarly writing.

Across these domains, the integration of AIGC is not merely a matter of tooling but a reconfiguration of labor, responsibility, and epistemology. Organizations deploying AIGC must consider sector-specific risks, user expectations, and governance requirements. The balance between augmentation and automation remains a moving target, contingent on task complexity, regulatory maturity, and public trust.

5. Ethical Challenges in AIGC Development and Deployment

As AI-generated content (AIGC) becomes deeply integrated into public discourse, creative production, and institutional processes, ethical considerations have emerged as a central concern in both academic research and industrial deployment. The ability of AIGC systems to mimic human communication and generate persuasive, authoritative content at scale creates unprecedented opportunities—and equally unprecedented risks. These risks are amplified by the opacity of model behavior, the scale of deployment, and the complexity of human-AI interaction. This section outlines the key ethical challenges associated with AIGC, focusing on content authenticity, bias and discrimination, accountability, value alignment, and the erosion of human agency.

A primary ethical issue concerns the authenticity and traceability of generated content. AIGC systems produce content that is increasingly indistinguishable from that created by humans. While this technological achievement is impressive, it also enables malicious uses such as disinformation campaigns, fake news dissemination, and synthetic identity fabrication. Deepfake videos, AI-written propaganda, and automated astroturfing efforts have already been documented in political, financial, and social contexts. Without robust detection tools, distinguishing AI-generated content from authentic human expression becomes increasingly difficult, undermining trust in digital communications. Research into watermarking, content provenance metadata, and model fingerprinting offers partial solutions, but none are yet universally deployed or immune to adversarial evasion [21].

Another persistent challenge is algorithmic bias. AIGC models are trained on large corpora of human-generated data, which often encode societal biases related to race, gender, nationality, and ideology. As a result, generated outputs can inadvertently perpetuate stereotypes, marginalize underrepresented groups, or reflect hegemonic worldviews. Bias in AIGC is not only a technical flaw but a social justice issue, as outputs may reinforce systemic inequities or exclude minority perspectives. For example, text-to-image models may associate certain professions with specific genders, or language models may produce different tone and sentiment for the same prompt depending on the demographic context. Bias mitigation requires a combination of diverse training data, fairness-aware pretraining objectives, and rigorous evaluation metrics—but these solutions remain nascent and contested [22].

Responsibility and accountability form another central ethical dilemma. When an AIGC system produces harmful, defamatory, or misleading content, it is often unclear who bears responsibility: the model developer, the deployer, the prompt engineer, or the end user. The distributed nature of AIGC pipelines complicates liability assignment and weakens traditional accountability frameworks. Moreover, open-source releases of powerful generative models enable downstream actors to modify or misuse them in ways that evade governance mechanisms. Calls for algorithmic transparency, third-party auditing, and model usage licensing have grown, but the technical and legal tools to enforce such accountability remain underdeveloped. Some researchers propose “model cards” and “system cards” as structured disclosures of model capabilities, risks, and limitations, but adoption is inconsistent [23].

Value alignment poses a further ethical challenge. AIGC systems do not possess intrinsic goals, values, or ethical intuitions. Their behavior is shaped by training data, optimization objectives, and human feedback signals, which may not reflect normative societal values or context-sensitive reasoning. Efforts to align generative models with human intentions—such as reinforcement learning with human feedback (RLHF)—have made progress, but alignment remains fragile. Misaligned systems can generate harmful instructions, inappropriate content, or misleading claims, particularly when exposed to ambiguous prompts or adversarial users. Philosophical concerns also arise: whose values are encoded, whose culture is amplified, and who defines what constitutes “acceptable” output? These questions have no simple answer and require participatory design processes involving diverse stakeholders [24].

The final and perhaps most subtle ethical concern is the erosion of human agency. As AIGC systems become more persuasive, interactive, and context-aware, users may increasingly defer to AI outputs without critical engagement. In educational, creative, and professional settings, this may reduce skill development, displace human judgment, or encourage epistemic complacency. Moreover, overreliance on AIGC can alter the relationship between authorship and ownership, diminishing individual expression in favor of statistically plausible but impersonal outputs. In collective contexts, AIGC may influence group decisions, moderate public opinion, or shape cultural narratives in invisible ways. Preserving human agency requires designing systems that support collaboration rather than substitution, transparency rather than opacity, and empowerment rather than automation [25].

In response to these ethical challenges, a growing ecosystem of governance efforts has emerged. Institutions such as the OECD, IEEE, EU AI Act Committee, and UNESCO have proposed guidelines for trustworthy AI, emphasizing principles such as transparency, fairness, accountability, and human oversight. Industry-led initiatives—such as the Partnership on AI and OpenAI’s system cards—aim to set voluntary standards and best practices. However, the rapid pace of AIGC development outstrips the implementation of ethical safeguards. Proactive engagement, interdisciplinary research, and policy coordination will be essential to bridge this gap.

6. Mitigation Strategies and Governance Frameworks

Given the multifaceted ethical, technical, and societal risks posed by AI-generated content (AIGC), researchers, policymakers, and industry practitioners have converged on a growing suite of mitigation strategies and governance frameworks. These efforts span technical defenses, human-centered design principles, organizational protocols, and regulatory initiatives. Their shared objective is to ensure that AIGC systems operate transparently, fairly, and safely, while preserving innovation and public trust. This section provides an integrative review of current best practices, emerging techniques, and institutional responses to mitigate AIGC risks.

A leading class of mitigation methods targets content authenticity and traceability. Digital watermarking—embedding imperceptible identifiers into generated outputs—has gained traction as a tool for post-hoc detection and provenance verification. Methods such as robust token-level watermarking in text, imperceptible pixel perturbation in images, or metadata embedding in video frames aim to mark AI outputs without degrading their quality. Google DeepMind's SynthID and OpenAI's watermarking research represent prominent examples of practical deployment. However, watermarking remains susceptible to removal or obfuscation through paraphrasing, cropping, or reformatting, highlighting the need for complementary detection methods [26].

In parallel, AI content detectors that leverage statistical cues (e.g., entropy, repetition, burstiness) or deep classification networks have been proposed to identify machine-generated outputs. While these tools are valuable for moderation and auditing, their reliability is constrained by adversarial prompt injection, distributional shift, and false positives against non-native or stylized human writing. Hybrid approaches combining watermark verification, style classifiers, and retrieval-based provenance matching offer a more resilient strategy, particularly in high-risk domains like education, finance, and scientific publishing [27].

Prompt control and content filtering are core to the real-time alignment of AIGC systems. Many production deployments incorporate input sanitization (e.g., prompt blacklists, syntactic validation) and output filtering using neural classifiers trained to detect hate speech, misinformation, sexual content, and violent narratives. Safety-tuned models also use reinforcement learning with human feedback (RLHF) to shape their generative behavior toward helpful, harmless, and honest outputs. Tools like OpenAI's moderation API, Anthropic's Constitutional AI, and DeepMind's Sparrow architecture introduce rule-based or policy-aware reasoning to enhance model reliability. Nevertheless, over-filtering may result in excessive censorship, while under-filtering risks exposing users to harm. Designing safety systems that are configurable, culturally adaptive, and explainable remains an open challenge [28].

At the systems level, sandboxing and access control mechanisms are used to contain the potential misuse of powerful generative models. Organizations may impose tiered access, limiting advanced capabilities to authenticated, rate-limited, or enterprise-level users. Audit logging, usage caps, and real-time anomaly detection enable administrators to monitor interactions and intervene if necessary. For open-source models, developers increasingly implement usage agreements, red-teaming protocols, and usage restrictions on downstream applications. Safety testing benchmarks and adversarial evaluations (e.g., using "jailbreak prompts") are critical for validating the robustness of such safeguards prior to deployment [29].

Beyond technical mitigation, institutional governance frameworks are evolving rapidly. At the policy level, the European Union's AI Act introduces tiered regulation for general-purpose AI and mandates transparency for synthetic content. It requires providers to disclose training data provenance, implement post-deployment monitoring, and allow opt-out from AI-based personalization. The U.S. White House has published an AI Bill of Rights, emphasizing user agency, data protection, and explainability. China's Cyberspace Administration has enacted content-specific AIGC guidelines emphasizing "truthfulness, fairness, and positive values." These frameworks reflect both local sociopolitical contexts and shared concerns around AI accountability and sovereignty [30].

Industry bodies such as the Partnership on AI, the OECD's AI Principles, and the IEEE Global Initiative on Ethically Aligned Design have developed normative guidelines and auditing tools. These initiatives call for transparency disclosures (e.g., model cards, system cards), risk impact assessments, and participatory oversight involving diverse stakeholders. Technical standards efforts (e.g., ISO/IEC JTC 1/SC 42) are working toward formalizing evaluation metrics, safety thresholds, and deployment protocols. Yet, harmonizing these global efforts into enforceable, interoperable standards remains a major governance challenge.

Finally, public education and media literacy programs are emerging as essential non-technical countermeasures. Empowering users to recognize, question, and verify AI-generated content is critical in environments where synthetic media may influence voting behavior, market decisions, or social perceptions. Civic organizations and academic institutions have begun incorporating AIGC awareness into digital literacy curricula. Toolkits for journalists, educators, and civil society actors are also being developed to promote responsible AIGC engagement across sectors.

In conclusion, effective mitigation of AIGC risks requires a multi-layered, interdisciplinary approach that integrates technical defenses, institutional controls, and societal adaptation. While no single solution is sufficient, the convergence of watermarking, safety-tuning, governance protocols, and transparency mechanisms offers a promising foundation for the safe and responsible advancement of generative AI technologies.

7. Conclusion

The rise of AI-generated content (AIGC) marks a pivotal transition in the evolution of artificial intelligence, signaling the shift from pattern recognition to autonomous content creation. Enabled by advancements in large-scale generative models—such as transformers, diffusion architectures, and reinforcement learning with human feedback—AIGC systems have demonstrated remarkable capabilities in producing natural language, imagery, software code, and multi-modal content. Their integration into critical domains such as journalism, education, legal practice, and creative industries has begun to reshape workflows, augment human expertise, and democratize access to knowledge and expression.

This paper has presented a comprehensive survey of AIGC from both technical and ethical perspectives. We examined the architectural foundations of modern generative systems, including their model design, training regimes, and deployment pipelines. We analyzed the system-level considerations involved in inference optimization, safety control, and multi-user interface integration. The discussion extended to domain-specific applications, illustrating the scope and versatility of AIGC across sectors while acknowledging the emerging tensions between automation and human agency. Furthermore, we systematically explored the ethical challenges associated with synthetic content—ranging from misinformation and algorithmic bias to value misalignment and accountability gaps.

To address these risks, a growing body of mitigation strategies has been developed, including digital watermarking, content filtering, sandboxed deployment, and alignment training. In parallel, regulatory frameworks and policy guidelines—such as the EU AI Act and the U.S. AI Bill of Rights—are beginning to define the contours of acceptable AIGC deployment. Nevertheless, technical, institutional, and normative gaps remain. Mitigation efforts are fragmented, and current governance mechanisms struggle to keep pace with the rate of innovation and global diffusion of generative technologies.

Looking forward, the development of safe and reliable AIGC systems will require coordinated progress on multiple fronts. Technically, future research must enhance model transparency, reduce hallucination, enable context-aware reasoning, and formalize mechanisms for controllable generation. Institutionally, developers and deployers must commit to documentation standards, impact auditing, and participatory governance. Societally, public education and digital literacy will be crucial in preparing users to navigate environments increasingly populated by synthetic content.

More fundamentally, the future of AIGC raises open questions about the boundaries of authorship, the nature of creativity, and the role of artificial intelligence in collective knowledge production. These are not merely technical challenges but epistemological and ethical ones that require interdisciplinary engagement. As such, the AIGC research agenda must move beyond performance metrics and benchmark scores to include considerations of justice, accountability, and cultural pluralism.

In conclusion, AI-generated content represents both an extraordinary opportunity and a profound responsibility. Its advancement offers the potential to augment human imagination, accelerate discovery, and expand access to digital expression. But realizing these benefits requires deliberate design, inclusive policy, and continuous reflection on the societal values we encode into the machines we build. This survey aims to support that endeavor by providing a structured foundation for future inquiry, development, and regulation in this rapidly evolving domain.

References

[1] B. McGuffie and A. Newhouse, “The Rise of AI-Assisted Journalism: Promise and Peril,” J. Media Technol. Ethics, vol. 5, no. 2, pp. 101–113, 2024.

[2] M. Zhai et al., “AI Tutors in the Classroom: An Evaluation of Pedagogical Impact,” IEEE Trans. Learn. Technol., vol. 16, no. 1, pp. 88–98, 2025.

[3] D. Katz and M. Bommarito, “ Generative AI in the Legal Profession: Applications, Risks, and Recommendations,” Harvard J. Law Tech., vol. 36, no. 3, 2024.

[4] L. Elkin-Koren, “Copyright Law and the AI Creator: Who Owns AI-Generated Art?,” Columbia J. Law Arts, vol. 47, no. 1, pp. 1–26, 2024.

[5] A. Vaswani et al., “Attention Is All You Need,” in Proc. NeurIPS, 2017.

[6] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.

[7] R. Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models,” in Proc. CVPR, 2022.

[8] J. Ho et al., “Classifier-Free Guidance for Diffusion Models,” in Proc. NeurIPS, 2022.

[9] L. Ouyang et al., “Training Language Models to Follow Instructions with Human Feedback,” arXiv preprint arXiv:2203.02155, 2022.

[10]S. Izacard et al.,“Unifying Retrieval-Augmented Generation and RAG,”Trans. Mach. Learn. Res., 2023.

[11]H. Bai et al.,“LoRA: Low-Rank Adaptation of Large Language Models,” arXiv preprint arXiv:2106.09685, 2023.

[12]T. Gao et al.,“Scaling Data Pipelines for Foundation Model Training: Lessons from GPT-4 and Beyond,”arXiv preprint arXiv:2401.05476, 2024.

[13]J. Lin et al.,“vLLM: Fast Inference and Serving for Large Language Models,”arXiv preprint arXiv:2309.06180, 2023.

[14]OpenAI,“System Card for GPT-4,”OpenAI Documentation, 2023. [Online]. Available: https://openai.com

[15]J. Yao et al.,“ReAct: Synergizing Reasoning and Acting in Language Models,”arXiv preprint arXiv:2210.03629, 2023.

[16]B. McGuffie and A. Newhouse,“The Rise of AI-Assisted Journalism: Promise and Peril,” J. Media Technol. Ethics, vol. 5, no. 2, pp. 101–113, 2024.

[17]M. Zhai et al.,“AI Tutors in the Classroom: An Evaluation of Pedagogical Impact,”IEEE Trans. Learn. Technol., vol. 16, no. 1, pp. 88–98, 2025.

[18]D. Katz and M. Bommarito,“Generative AI in the Legal Profession: Applications, Risks, and Recommendations,”Harvard J. Law Tech., vol. 36, no. 3, 2024.

[19]L. Elkin-Koren,“Copyright Law and the AI Creator: Who Owns AI-Generated Art?,”Columbia J. Law Arts, vol. 47, no. 1, pp. 1–26, 2024.

[20]H. Pearce et al.,“Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions,” IEEE Symp. Secur. Privacy, pp. 754–768, 2023.

[21]R. Krishna et al.,“Detecting AI-Generated Content: A Survey of Techniques and Limitations,” ACM Comput. Surv., vol. 56, no. 2, pp. 1–39, 2024.

[22]A. Bender and E. Friedman,“Bias in Text and Image Generators: A Cross-Modal Review,”IEEE Trans. Technol. Soc., vol. 4, no. 1, pp. 51–67, 2024.

[23]M. Mitchell et al.,“Model Cards for Model Reporting,”in Proc. FAT, 2019, pp. 220–229.

[24]J. Steinhardt,“Alignment Research in the Era of Generative Models,” AI Ethics J., vol. 3, no. 4, pp. 215–230, 2023.

[25]S. Gabriel et al.,“Human Agency and AI Systems: Designing for Meaningful Control,”Philos. Technol., vol. 36, no. 1, 2023.

[26]Y. Zhang et al.,“Robust Watermarking for Large Language Models via Context-Aware Token Perturbation,” arXiv preprint arXiv:2306.04094, 2023.

[27]C. Carlini et al.,“Testing for Content Memorization in Generative Models,” Proc. IEEE Symp. Secur. Privacy, pp. 1021–1038, 2023.

[28]J. Bai et al.,“Constitutional AI: Harmlessness from AI Feedback,”arXiv preprint arXiv:2212.08073, 2022.

[29]R. Ganguli et al., “Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned,” arXiv preprint arXiv:2305.17564, 2023.

[30]European Commission, “Artificial Intelligence Act: Proposal for a Regulation,” Brussels, 2023. [Online]. Available: https://ec.europa.eu