IRIS
Information Resilience
& Integrity Symposium

RESEARCH PAPER

Can We Trust AI? Understanding Moral Reasoning Beyond GenAI Utilisation

PANEL 1 | Deepfakes for Financial Fraud

Safer Internet Lab
Led by CSIS in partnership with Google

Centre for Strategic and International Studies

Google

Universitas Gadjah Mada
Fakultas Ilmu Sosial dan Ilmu Politik

Center for Digital Society

Dian Kartika Rahajeng, Ph.D

Dian Kartika Rahajeng, Ph.D is an Assistant Professor at Universitas Gadjah Mada, Indonesia, with expertise in Islamic finance, fraud investigation, forensic accounting, auditing, and governance issues. She has co-authored books and articles on corporate governance concepts and their implementation in traditional and non-traditional organizations. Dian's research interests include ethics and moral economy in traditional and non- traditional organizations, particularly in Islamic entities. She is an expert witness in high-profile crime cases, specializing in corruption and money laundering schemes within fraudulent corporations. Her commitment to academic and professional excellence has earned her numerous prestigious awards. She can be contacted at dkrahajeng@ugm.ac.id

This paper is circulated for discussion and feedback. The views expressed are solely those of the author(s) and do not represent an official position of SAIL, CSIS, Google, CfDS, Faculty of Social and Political Sciences UGM or any other organization. The author(s) welcome comments on this version and invite you to contact them directly with any feedback or questions.

Can We Trust AI? Understanding Moral Reasoning Beyond GenAI Utilisation

Dian Kartika Rahajeng, Ph.D

This study explores the role of generative artificial intelligence (GenAI) in ethical decision-making frameworks, focusing on cognitive morality and its relationship with deepfake technologies. The research uses qualitative approach, starting with an extensive systematic literature review, documentary analysis, and in-depth interviews with 21 entry-level managers across various global organizations. The findings suggest that while GenAI can enhance decision-making processes, it lacks the autonomous moral reasoning capacity necessary for generating ethical agencies. Therefore, cognitive morality and accountability must be human-driven and grounded in advanced stages of the agents' cognitive moral development. The study provides real-world examples, such as GenAI-generated deepfakes impersonating executives and authoring fraudulent financial transactions, exposing vulnerabilities within trust paradigms. The research contributes to the discourse on AI and ethics by synthesizing cognitive moral theory with actionable organizational insights, addressing a gap in the literature on GenAI's ethical boundaries. It emphasizes the need for robust ethical safeguards and human oversight in AI systems and calls for organizations to strengthen governance structures, implement rigorous verification protocols, and cultivate advanced digital literacy to mitigate risks associated with AI misuse.

Keywords: GenAI, Ethical Decision-Making, Moral Reasoning, Trust, Ethics, Implications

Introduction

The pervasive influence of artificial intelligence (AI) has dramatically reshaped societal structures and professional practices across sectors. In recent years, the beginning of Generative AI (GenAI) and models like ChatGPT have instigated a significant transformation, prompting both immense excitement and considerable anxiety regarding their far-reaching implications (Gordijn & Have, 2023; Larsen & Narayan, 2023). While these technologies promise unprecedented opportunities for innovation and efficiency, they simultaneously introduce a complex array of ethical, social, and practical challenges that demand critical attention (Susarla et al., 2023).

The growing literature on GenAI extensively explores its capabilities and applications. However, a gap remains large in the depth of discourse surrounding the ethical dimensions of GenAI adoption. This paper seeks to bridge this critical gap by enriching the preliminary debates by adding an extensive literature review of the alignment between ethics and GenAI and enriching it with qualitative studies of entry-level manager acceptance towards GenAI daily utilization in their decision-making process. Our central tenet is that irrespective of the emerging ethical perspectives on AI, the indispensable roles of the actor's (human) moral perception and critical, logical reasoning remain paramount. These cognitive faculties constitute the core of human existence and ethical deliberation, advocating for an inherently human-centric approach to AI development and integration, rather than an objective-centric one. This necessitates a delicate balance, wherein ethical values, intrinsically human in origin, are consciously embedded and perpetually reinforced within the design and operation of Large Language Models (LLMs).

The urgency of this discourse is underscored by the rapid proliferation of deepfake technology. Deepfakes, which are artificially manipulated audio and visual content generated by AI, are becoming increasingly sophisticated and accessible (Sherman, 2024). This accessibility has enabled criminals to orchestrate elaborate financial fraud schemes, resulting in substantial monetary losses globally (Sherman, 2024; Gupta et al., 2025). The ability of deepfakes to convincingly impersonate individuals, including high-ranking executives, to authorize fraudulent transactions highlights the critical vulnerabilities within existing trust frameworks and emphasizes the imperative for enhanced human oversight and robust ethical guidelines in AI systems.

This paper is structured to systematically address these complex interdependencies. Section 1 will further delineate the current research landscape and underscore the necessity of this study in enriching the dialogue on ethical GenAI adaptation. Section 2 will provide a comprehensive overview of the latest advancements in GenAI. Section 3 will detail the qualitative perspectives employed in this research. Section 4 presents a thorough discussion of the findings, implications, and concrete examples of deepfake financial fraud critically examining the interplay of ethical values and GenAI's applicability. Finally, Section 5 will offer a constructive conclusion, delineate the study's limitations, and provide valuable insights for future research.

Literature Gap: A Need for Deeper Ethical Scrutiny of GenAI

While the capabilities of GenAI have captivated researchers and practitioners alike, leading to a prolific volume of literature on its technical advancements and diverse applications (Lachheb et al., 2024), there is a growing acknowledgment that the ethical implications of GenAI's widespread adaptation warrant far more rigorous and nuanced examination. Existing reviews often focus on trend mapping or thematic categorization without a critical synthesis of how AI in higher education, for instance, is framed across disciplines or how broader social, historical, or epistemic dimensions influence this framing (Deng et al., 2024). This gap extends to the broader application of GenAI across industries, including global financial systems.

The rapid evolution of GenAI challenges traditional ethical frameworks, necessitating a re-evaluation of concepts such as authorship, responsibility, and trust (Zohny & McMillan, 2023). The ability of GenAI to produce content that is virtually indistinguishable from human-generated text, as demonstrated by GPT-3, raises profound questions about authenticity and the potential for automated mass manipulation (Illia et al., 2023). This phenomenon directly impacts the integrity of information and underscores the urgent need for comprehensive research and countermeasures to mitigate its adverse effects (Muhly et al., 2025).

Moreover, the convergence of GenAI with deepfake technologies presents a particularly alarming challenge to information security and financial stability. As deepfakes become more realistic and accessible, their malicious use in financial fraud, market manipulation, and corporate espionage is escalating (Gupta et al., 2025; Bateman, 2020). Despite the documented cases of deepfakes being used for fraud and extortion, the financial threat from synthetic media is often underestimated (Bateman, 2020). There remains a significant lack of research that integrates cognitive moral theory with practical organizational insights, specifically addressing the moral boundaries of GenAI as an ethical tool within such high-stakes contexts. This paper aims to fill this critical void, contributing to a more robust and human-centric understanding of AI ethics.

Latest Advances in Generative AI and Deepfake Technology

The rapid advancements in Artificial Intelligence (AI) have ushered in a new era of generative capabilities, profoundly impacting various domains from content creation to cybersecurity. Central to this evolution are GenAI models, particularly LLMs like ChatGPT, and the sophisticated deepfake technology that leverages these advancements. This section delineates the latest developments in these areas, highlighting their underlying mechanisms and the increasing sophistication of deepfake generation and detection.

Artificial intelligence, broadly understood as a system's capacity to interpret external data, learn from it, and apply that knowledge to achieve specific goals through flexible adaptation (Kaplan & Haenlein, 2019; Laine et al., 2025), has continuously pushed computational boundaries for over seven decades. From early conceptualizations, such as McCulloch and Pitts's neuron-inspired computer model in 1943, and Turing's 1950 introduction of the Turing test to gauge AI intelligence, the field has seen consistent expansion (Muthukrishnan et al., 2020; Kaplan & Haenlein, 2019; Laine et al., 2025). This progression is exemplified by the evolution from the first AI chatbot, ELIZA (Weizenbaum, 1966), to today's highly sophisticated conversational systems like Apple's Siri and Amazon's Alexa (Adamopoulou & Moussiades, 2020; Fui-Hoon Nah et al., 2023; Berente et al., 2021; Laine et al., 2025).

The field of AI experienced a significant transformation with the emergence of generative models, marking a new phase in AI's capacity to process and create complex content. This period saw the development of advanced generative AI approaches, including hidden Markov models and Gaussian mixture models (Cao et al., 2023; Gupta et al., 2023). Generative AI distinguishes itself by leveraging deep learning models to produce human-like content, such as images or text, in response to complex and varied prompts like natural language instructions or questions (Lim et al., 2023; Laine et al., 2025).

Unlike traditional expert systems that primarily analyze or act on existing data, GenAI is designed to generate novel content, capable of producing textual, visual, and auditory material with minimal human intervention (Gozalo-Brizuela & Garrido-Merchan, 2023; Longoni et al., 2022). Its core function is to generate new data, rather than making decisions based on existing information (Jovanovic & Campbell, 2022). The outputs of GenAI systems are inherently diverse, open-ended, and often unpredictable (Dwivedi et al., 2023; Mökander et al., 2023), requiring relatively minimal input.

These systems are trained on vast and diverse datasets sourced from the internet (Korzynski et al., 2023; Mökander et al., 2023; Stokel-Walker & Van Noorden, 2023), allowing them to exhibit creativity and produce fresh content that goes beyond rule-based decision-making (Pavlik, 2023; Zhuo et al., 2023). Currently, four prominent GenAI techniques include generative adversarial networks (GANs), Generative Pre-trained Transformers (GPTs), generative diffusion models (GDMs), and geometric deep learning (Jovanovic & Campbell, 2022). GenAI systems generate new data, generating diverse, open-ended, and unpredictable outputs that require minimal input. These systems are trained on vast internet datasets, allowing them to exhibit creativity and produce fresh content beyond rule-based decision-making. These systems exhibit creativity and produce fresh content beyond rule-based decision-making.

A pivotal leap in AI's evolution was the advent of Large Language Models (LLMs), which profoundly reshaped natural language processing (NLP) and conversational AI. LLMs are characterized by their unprecedented scale and complexity, being trained on massive datasets containing billions of parameters and requiring substantial computational resources. They possess versatile capabilities, adapting responses in real-time based on user input and evolving contexts, though this reliance on extensive training data also raises significant concerns about data privacy and security (Meskó & Topol, 2023). This advancement led to the development of sophisticated AI chatbots such as AlphaGo in 2015 and, more recently, Google Bard and OpenAI's ChatGPT in 2022 (Fui-Hoon Nah et al., 2023; Meskó & Topol, 2023). These AI chatbots are a subset of conversational AI, which utilizes machine learning and NLP to understand and respond to user input in natural language across various communication channels (Ruane & Birhane, 2019). While conversational AI typically relies on predefined responses, GenAI models can generate entirely new content, distinguishing them conceptually (Lim et al., 2023).

ChatGPT, launched in 2022 and rapidly becoming one of the fastest-growing consumer applications, epitomizes these LLM advancements (Chatterjee & Dethlefs, 2023; Van Slyke et al., 2023). Based on advanced GPT models, it excels at generating nuanced, human-like conversations (Lim et al., 2023). However, ChatGPT also highlights LLM limitations, including instances of providing incorrect answers, referencing non-existent scientific studies, producing plausible but inaccurate content, and exhibiting biases from its training data (Gordijn & Have, 2023; Meskó & Topol, 2023; Thorp, 2023). Although equipped with safeguards against sensitive or illegal topics, users have reportedly found ways to bypass them (Sun et al., 2024). As AI continues to integrate into various sectors, addressing these challenges becomes crucial to ensure ethical development and positive societal contributions.

Deepfake technology, a term coined in 2017, represents a significant advancement in synthetic media creation, enabling the manipulation and generation of highly realistic audio-visual content. Recent advancements have allowed models to create convincing deepfakes with minimal input, making the technology more accessible and, consequently, more dangerous (Mitra et al., 2021). Deepfake technology, often associated with illicit activities, has a dual nature and offers numerous benefits across various sectors. In the entertainment industry, it revolutionizes film production by enabling the de-aging of actors, the recreation of deceased performers, and the enhancement of visual effects.

In video games, deepfakes create more realistic characters and animations, while in education and training, they generate engaging and interactive learning experiences. In medical training, deepfakes simulate patient interactions for more realistic scenarios. In business communication and marketing, deepfakes help break down language barriers by translating and lip-syncing speeches for global audiences. Virtual spokespersons offer consistent brand representation, and enhanced video conferencing capabilities improve remote interactions. In art and accessibility, deepfakes are used as a new medium for surreal scenes, pushing the boundaries of visual art. They can also generate sign language interpretations or audio descriptions for visual media, making content more inclusive and accessible.

Deepfakes, despite their positive applications, pose significant threats when used maliciously. They can be used in political disinformation campaigns, creating false narratives about public figures or events that can sway public opinion and undermine democratic processes. The "liar's dividend" phenomenon allows individuals to discredit genuine evidence by falsely claiming it is fabricated. In personal and social contexts, deepfakes are most commonly misused for non-consensual pornography, cyberbullying, and online harassment, leading to emotional distress and reputational damage. Technology's ability to create convincing fake videos and audio opens new avenues for identity theft and financial fraud. Deepfake voice phishing (vishing) uses cloned voices to impersonate trusted individuals, exploiting professional or personal relationships to induce fraudulent transfers. Deepfakes can manipulate financial markets by fabricating statements from executives, causing stock price fluctuations. In corporate espionage, they can be used to impersonate individuals to gain unauthorized access to sensitive information or influence critical business decisions. Deepfakes represent a new frontier in the spread of fake news, making it increasingly challenging for journalists and fact-checkers to combat misinformation due to the highly convincing nature of fabricated video evidence.

Deepfake generation has led to the development of various detection techniques, each with its strengths and weaknesses. AI-based methods, such as Convolutional Neural Networks (CNNs), are accurate but require significant computational resources and can overfit training data (El-gayar et al., 2024). Forensic analysis provides detailed insights into manipulation techniques, but is time-consuming and less effective for real-time detection. As generative models become more sophisticated, detection methods must adapt to identify subtle inconsistencies and unique "fingerprints."

Data and Methodology

This study employs a qualitative approach, conducting semi-structured interviews and focus group discussions with selected entry-level managers in global organizations in Indonesia, France, Switzerland, the Netherlands, and Germany. A comprehensive documentary review was conducted to gather and synthesize information from a broad range of academic literature and media coverage related to GenAI, deepfakes, and financial fraud.

The study involved 21 early-to-mid-career professionals aged 25-32 years, focusing on their perceptions of GenAI and its ethical standing. The majority, 61.9%, were employed in international organizations, while the remaining 38.1% worked in national or multinational organizations. The demographic consisted of 71.4% females and 28.6% males.

Table 1. Participation list

No	Participant Code	Gender	Organization country-based
1	FR1	F	France
2	FR2	F	France
3	FR3	M	France

4	FR4	F	France
5	FR5	M	France
6	FR6	M	France
7	FR7	F	France
8	GER1	F	Germany
9	GER2	F	Germany
10	GER3	F	Germany
11	GER4	M	Germany
12	INA1	M	Indonesia
13	INA2	F	Indonesia
14	INA3	F	Indonesia
15	INA4	F	Indonesia
16	INA5	F	Indonesia
17	INA6	F	Indonesia
18	INA7	M	Indonesia
19	INA8	F	Indonesia
20	NED1	M	Netherlands
21	SWZ1	F	Switzerland

Source: Author's own work

Each participant was chosen based on a preliminary review of previous focus group discussions. The researcher's personal background, geographic setting, and professional networks all influence the recruiting process. During focus group discussions in class settings, following up with semi-structured interviews, participants were asked about the integration of AI technologies into their professional activities, their intensity, and perceptions of AI's ethical capacity, and whether AI systems can possess ethical standing in their utilization. The semi-structured interview and focus group discussion duration is 45 minutes to 90 minutes.

The data was transcribed and subjected to thematic analysis, which involved familiarizing with the data through repeated readings. Initial codes were generated and grouped into themes related to AI ethics, trust, moral agent, and fraud. These themes were reviewed and refined to ensure they accurately represented the data. The themes were presented with illustrative quotes from the interview data, integrated with conceptual and documentary analyses. The study's findings are enhanced by the triangulation of conceptual analysis, documentary review, and qualitative interviews.

Qualitative research distinguishes itself from quantitative methodologies by prioritizing an in-depth exploration of human experiences, thereby providing nuanced insights into individual perceptions and assigned meanings, rather than focusing on statistical generalization. In this paper, data collected from 21 participants provided extensive empirical material, facilitating a comprehensive understanding of their viewpoints. Although the findings are primarily derived from subjective perceptions, their validity is significantly strengthened through the triangulation of data with documentary analysis. This multi-methodological approach ensured the findings were rigorously corroborated and iteratively compared against existing academic literature and empirical studies, achieving saturation where no new conceptual themes emerged. Consequently, the conclusions drawn are considered robust and adequately supported. Nevertheless, incorporating a quantitative approach in future research could further enhance discourse by examining the prevalence and generalizability of these qualitative insights across a larger population.

Discussion

The results of this study reveal a clear divergence in perspectives regarding the ethical capabilities of GenAI. While approximately half of the respondents acknowledged the increasing utility of GenAI, particularly in critical sectors such as medical diagnostics and autonomous military operations, viewing AI as a growing influence in ethical decision-making due to its expanding role in life-altering contexts, the other half expressed significant reservations. This latter group strongly resisted attributing any form of ethical agency or inherent rights to AI systems, asserting that GenAI, despite its functional effectiveness, remains a tool created and controlled by human beings (see quotation NED1 below).

> “Robots are actually tools created and programmed by human beings, and their autonomy is only artificial.” (NED1)

Consequently, the ultimate responsibility for any decisions or actions generated by AI, whether intended or unintended, firmly resides with its human creators, designers, and operators, not the technology itself.

> “Granting rights to robots/AI, a set of instrumental rights [moral consciousness] can be established. These will not consider the robot as a moral agent but regulate other agents' behaviors in respect of its appropriate and ethical incorporation into society. This may be things like protections against abuse, destruction, or liability for programming bias or potentially damaging uses.” (NED1)

This perspective is grounded in the belief that GenAI lacks independent consciousness, genuine moral reasoning, and the capacity for accountability, rendering it unsuitable for the designation of a moral agent. Across both viewpoints, a pervasive theme emerged: the critical importance of maintaining human-centric accountability frameworks. Even among those who advocated for broader AI implementation, there was a universal consensus that ethical reasoning and ultimate responsibility must remain deeply rooted in human cognitive development and moral judgment. These findings collectively reinforce the argument that while GenAI can significantly support and even simulate aspects of ethical decision-making, it cannot authentically embody ethical values, rights, or responsibilities in the human sense.

The rapid advancement of GenAI presents a profound ethical challenge, particularly when its capabilities are intertwined with malicious applications such as deepfake financial scams. While AI offers immense potential for societal benefit, its design and deployment must be meticulously aligned with human values, ensuring that technological progress serves humanity rather than undermining its core principles (Groumpos, 2022; Coeckelbergh, 2020).

The Human-Centric Imperative

A central argument is the necessity of a human-centric approach to AI ethics. This study echoes Coeckelbergh (2020), who argues that moral cognition and accountability are intrinsically human attributes rooted in consciousness, empathy, and complex reasoning that AI currently lacks and may never fully replicate. As per Kohlberg's theory of cognitive moral development, genuine moral agency progresses through stages of conventional and post-conventional reasoning, where individuals internalize universal ethical principles and can make autonomous judgments that go beyond mere rule-following or self-interest.

GenAI, operating on algorithms and data, can mimic human outputs but cannot grasp the nuanced moral implications or the inherent values underpinning ethical decisions.

Interviewed managers (i.e., NED1) consistently emphasized the irreplaceable role of human judgment. NED1 stated, "While AI can provide vast amounts of information and even suggest solutions, the ultimate decision, especially in ethical dilemmas, must rest with a human." An AI cannot understand the 'why' behind our values. This sentiment underscores that ethical decision-making is not merely a computational task but a deeply human process involving intuition, lived experience, and an understanding of societal norms. SWZ1 added that "AI/robot could have the right not to be held responsible for malfunctions caused by poor programming or inadequate maintenance by the manufacturer. This would put the focus on the developers and operators and oblige them to maintain the highest [moral] standards."

The study argues that while AI may exhibit a "weak" form of moral agency due to the consequential nature of its actions, it fundamentally lacks the "stronger" form rooted in genuine moral reasoning, thereby emphasizing that ultimate accountability must remain human-driven. A critical aspect of this analysis is critiquing the alignment of ethics with GenAI by asserting that ethical frameworks must remain human-centric, prioritizing human values, critical logic, and sensory perception over purely objective, algorithm-driven outcomes. This perspective posits that ethical values are inherently human constructs that must inform and be continuously fed into LLMs, rather than being passively derived or emergent from AI processes (Coeckelbergh, 2020).

Utilizing Kohlberg's theory of moral development to analyze the limitations of AI in achieving higher-order moral reasoning (Velasquez, 2002), while capable of complex computations, GenAI cannot replicate the human capacity for post-conventional moral reasoning, which involves abstract principles and universal ethical considerations that transcend mere rule-following. This theoretical lens helps to articulate why cognitive morality and accountability must remain human prerogatives. These principles encompass beneficence, non-maleficence, autonomy, justice, explicability, fairness, reliability, safety, privacy, security, inclusiveness, transparency, and accountability.

While the ethical principles mentioned are laudable, their practical application in GenAI is uptight with challenges. Deepfake-like technology can offer beneficial accessibility features, for instance, by generating sign language interpretations or audio descriptions that make content accessible to individuals with impairments (Ceolin, 2023), thereby directly supporting the principle of inclusiveness. GenAI can also serve as a tool for creative augmentation, enabling artists to explore new mediums and push the boundaries of visual art (Wagner & Blewer, 2019). This enhances human creativity rather than replacing it, fostering a collaborative model between humans and AI.

Conversely, the unethical utilization of GenAI presents several critical challenges. A significant concern lies in bias and discrimination, where AI algorithms trained on biased historical data can perpetuate and amplify existing societal inequalities. Another critical area is privacy violations; GenAI-powered facial recognition technology, for example, can identify individuals and predict sensitive information, such as emotional states or even sexual preferences, without explicit consent (Coeckelbergh, 2020). This raises serious privacy concerns and can be exploited for surveillance, thereby violating personal autonomy. Automated mass manipulation and disinformation represent another profound ethical challenge, as GenAI can generate highly refined, human-like text at scale, leading to "fake agenda problems" and the proliferation of low-quality content (Illia et al., 2023).

Critical Reasoning: Why GenAI Cannot Be Fully Reliable

The limitations of GenAI in achieving true moral reasoning and full reliability stem from several core distinctions between artificial and human intelligence. GenAI operates based on statistical patterns and programmed objectives, devoid of consciousness, subjective experience, or genuine intent. It does not "understand" consequences in the human sense or possess a moral compass; its "ethical" behavior is merely a reflection of the ethical values encoded in its training data and algorithms, which are human-derived.

Human moral decisions are deeply embedded in social, cultural, and emotional contexts. GenAI, despite its advanced capabilities, cannot fully grasp these nuanced contexts or experience empathy. It cannot "feel" the suffering caused by its decisions or truly comprehend the complex web of human relationships that inform ethical choices. Fundamentally, GenAI functions as a tool, an extension of human intelligence. Its ethical outputs are only as robust as the ethical data and frameworks it is based.

If the input data is biased, incomplete, or reflects flawed human values, the GenAI's output will inevitably reflect these deficiencies, underscoring the continuous human responsibility in curating and refining the data that shapes AI. Furthermore, while developers may have good intentions, AI can produce unintended and harmful consequences, such as reinforcing bias or spreading hate speech.

The "many hands" problem in AI development makes assigning responsibility challenging, thereby highlighting the need for clear human accountability frameworks. As one interviewee put it, "Who is truly responsible when an AI makes a wrong decision? The programmer, the user, the data provider? It's always a human, ultimately", (FR3). Lastly, human moral reasoning possesses the unique ability to adapt to novel, unprecedented ethical dilemmas, drawing on abstract principles and a capacity for creative problem-solving. GenAI, reliant on trained data, may struggle to navigate genuinely new moral quandaries that fall outside its learned patterns, potentially leading to unpredictable and ethically problematic outcomes.

Deepfake Financial Fraud: A Consequence of Unchecked GenAI Power

The ethical shortcomings and inherent unreliability of GenAI are starkly exposed in the context of deepfake financial fraud. These scams represent a critical convergence of advanced generative capabilities and malicious intent, demonstrating the urgent need for robust human oversight and intervention. Deepfake technology allows criminals to create highly convincing fake audio and video of individuals, enabling sophisticated imposter scams (Sherman, 2024; Gupta et al., 2025). A notable example is the Hong Kong bank heist, where a finance worker was tricked into transferring $35 million after a deepfake voice cloned that of a company director (Sherman, 2024). Figure 1 shows the documentary analysis from media coverage of the deepfake scam in Hong Kong. It shows that many individuals are responsible both directly and indirectly for the incidents. Everyone has roles as enablers and imposters of scams, including deepfakes in financial fraud.

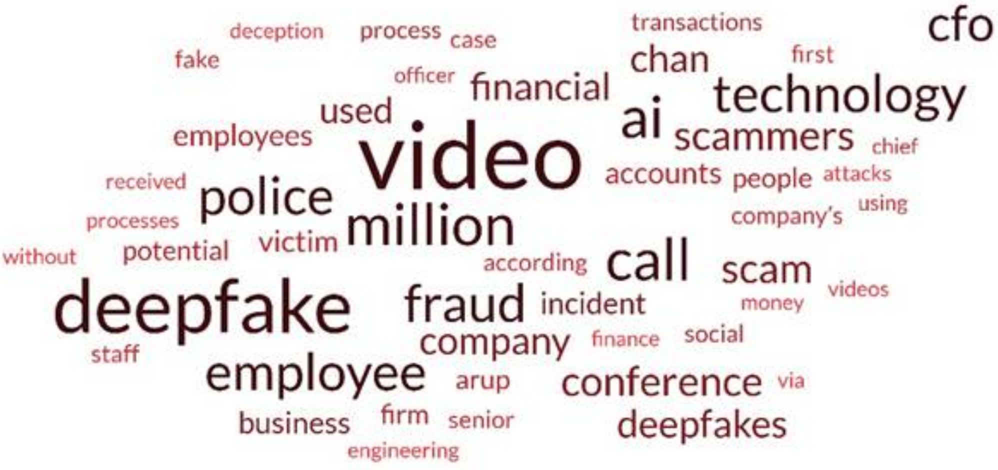

Figure 1. Word clouds of deepfake media coverage
Source: Author's work

In another incident, a hacker utilized an AI-generated voice deepfake during a phone call to obtain multi-factor authentication codes from an IT company's employees, impacting 27 cloud customers (Gupta, 2025). These "deepfake vishing" attacks exploit human trust and can be incredibly persuasive, effectively bypassing traditional security measures (Bateman, 2020). The psychological impact on victims is significant; individuals may attribute flaws in the cloned voice to a poor connection or be emotionally manipulated during high-pressure calls (Bateman, 2020).

Beyond individual targeting, deepfakes can be used to spread false information about companies, manipulating stock prices or triggering panic (Gupta et al., 2025; Bateman, 2020). Fabricated private remarks or fake news reports depicting public figures making damaging comments can lead to "pump and dump" or "short and distort" schemes (Bateman, 2020). While the overall threat to global financial stability may be low in mature economies, individual companies and emerging markets remain highly vulnerable (Bateman, 2020). The rapid spread of such misinformation, amplified by social media, can erode public trust in financial systems and institutions (Kraus, 2020; Gupta, 2025). The proliferation of deepfakes also fosters widespread skepticism toward all digital content, making it increasingly difficult for individuals and organizations to discern truth from falsehood (Ajder et al., 2019; Vaccari & Chadwick, 2020).

This "reality apathy" can undermine the credibility of media and institutions, potentially leading to increased conspiracy theories and societal division (Vaccari & Chadwick, 2020). Managers interviewed expressed a growing concern about employees' ability to critically evaluate digital communications, highlighting the paramount need for enhanced digital literacy. Despite advancements in deepfake detection techniques, the continuous evolution of generative AI creates an ongoing "arms race" (Mirsky & Lee, 2021). Current legal frameworks often lag behind technological advancements, complicating the prosecution of perpetrators and the resolution of issues related to consent, copyright, and defamation (Sherman, 2024; Chesney & Citron, 2019). The anonymity of deepfake creators further complicates enforcement efforts. The pervasive nature of deepfake financial fraud thus underscores that while GenAI offers transformative potential, its deployment without robust human ethical guidance and critical oversight poses substantial risks to financial integrity, individual trust, and societal stability. The emphasis must remain on human accountability and the continuous embedding of human ethical values into the AI development lifecycle.

Conclusion

This study has critically examined the intricate relationship between GenAI, ethical considerations, and the escalating threat of deepfake financial scams. Our central argument posits that regardless of the impressive capabilities of emerging AI technologies, the fundamental human capacities for sensory perception, critical logic, and moral reasoning remain indispensable. These attributes are not merely desirable additions but are foundational to human identity and ethical agency, necessitating an inherently human-centric approach to AI development and deployment. This paper argues that true ethical alignment in AI systems is achieved not through algorithmic objectivity alone, but through a continuous, balanced integration of human ethical values into the training and operational frameworks of Large Language Models (LLMs).

This research has underscored several key findings. While GenAI demonstrably augments decision-making processes, it fundamentally lacks the autonomous moral reasoning required to act as a genuine ethical agent. This absence of true moral cognition means that accountability must always remain human-driven, rooted in the advanced stages of moral development that AI cannot replicate.

The paper has underlined the dual nature of GenAI and deepfake technology, showcasing their potential for beneficial applications in many areas, while simultaneously highlighting their severe misuse in malicious schemes. Real-world examples of deepfake financial fraud, including sophisticated impersonation scams and market manipulation, vividly illustrate the vulnerabilities inherent in current trust frameworks when human oversight and ethical safeguards are insufficient. Ultimately, this study emphasizes that GenAI cannot be fully reliable as an ethical decision-maker due to its lack of consciousness, contextual understanding, empathy, and dependence on human-curated data. The "black box" problem and the challenge of unintended consequences further underscore the imperative for continuous human involvement and accountability throughout the AI lifecycle.

This paper significantly contributes to the evolving discourse on AI ethics by integrating cognitive moral theory with practical organizational perspectives, addressing a crucial gap in the literature concerning the moral boundaries of GenAI. The study provides valuable insights into AI ethics, but has limitations. The qualitative component, based on interviews with 21 entry-level managers, may not fully represent all organizational levels or global perspectives. The study also highlights the dynamic nature of AI and deepfake technologies, which may evolve rapidly.

Perceptions of AI ethics, trust, and moral responsibility can vary across individuals, cultures, and organizational contexts. The study's focus on GenAI and deepfake financial scams may limit its generalizability to other misuse contexts. Additionally, Kohlberg's theory of moral development could offer more nuanced insights into human moral reasoning in relation to AI. Future research should include senior leadership, AI developers, policymakers, and a wider geographical distribution.

References

A. Lachheb, J. Leung, V. Abramenka-Lachheb, et al., AI in higher education: A bibliometric analysis, synthesis, and a critique of research, *The Internet and Higher Education* (2024), https://doi.org/10.1016/j.iheduc.2025.101021

Adamopoulou, E., & Moussiades, L. (2020). *An Overview of Chatbot Technology* (pp. 373–383).

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The state of deepfakes: Landscape, threats, and impact. *Amsterdam: Deeptrace, 27*.

Bateman, J. (2020). *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios* (#7; Cyber Policy Initiative Working Paper Series). https://carnegie-production-assets.s3.amazonaws.com/static/files/Bateman_FinCyber_Deepfakes_final.pdf

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT.

Ceolin, F. (2023). Beyond deepfakes: The positive applications of AI-enhanced video synthesis [Accessed: 2024- 10- 02].

CFO Dive. (2024, May 17). Scammers siphon $25M from engineering firm Arup via AI deepfake 'CFO'. Retrieved from https://www.cfodive.com/news/scammers-siphon-25m-engineering-firm-arup-deepfake-cfo-ai/716501/

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107, 1753–1820.

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642.

El- gayar, M., Abouhawwash, M., Askar, S., & Sweidan, S. (2024). A novel approach for detecting deep fake videos using graph neural network. *Journal of Big Data*, 11(1), 22. Advance online publication. DOI: 10.1186/s40537- 024- 00884- y

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304.

FutureCIO. (2024, June 7). Hong Kong sees a 1000% increase in deep fake incidents. Retrieved from https://futurecio.tech/hong-kong-sees-a-1000-increase-in-deep-fake-incidents/

Globalnews.ca. (2024, February 5). Company out $35M after scammers stage video call with deepfake CFO, coworkers. Retrieved from https://globalnews.ca/news/10273167/deepfake-scam-cfo-coworkers-video-call-hong-kong-ai/

Gordijn, B., & Have, H. ten. (2023). ChatGPT: evolution or revolution? *Medicine, Health Care and Philosophy*, 26(1), 1–2.

Government Information Centre. (2024, June 26). LCQ9: Combating frauds involving deepfake. Retrieved from https://www.info.gov.hk/gia/general/202406/26/P2024062600192.htm

Gozalo-Brizuela, R., & Garrido-Merchan, E. C. (2023). ChatGPT is not all you need. A State-of-the-Art Review of large Generative AI models.

Groumpos, P. P., & PAPER, P. (2022). Ethical AI and Global Cultural Coherence: Issues and Challenges. *IFAC-PapersOnLine*, *55*(39), 358–363. https://doi.org/https://doi.org/10.1016/j.ifacol.2022.12.052

Gupta, G., Sailaja, B., Kovid, R. K., & Pandla, K. (2025). *Deepfakes and Their Impact on Business*. IGI Global Scientific Publishing.

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy.

Illia, L., Colleoni, E., & Zyglidopoulos, S. (2022). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, The Environment and Sustainability*, 1–414. https://doi.org/10.1111/beer.12479

Jovanovic, M., & Campbell, M. (2022). Generative Artificial Intelligence: Trends and Prospects. *Computer*, 55(10), 107–112.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25.

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories: analysis of ChatGPT. *Central European Management Journal*, 31(1), 3–13.

Kraus, S. (2020). The Growing Threat of Deepfakes and Synthetic Media in Financial Fraud.

Journal of Financial Crime, 27(3), 877–885.

Laine, J., Minkkinen, M., & Mäntymäki, M. (2025). Understanding the Ethics of Generative AI: Established and New Ethical Principles. Communications of the Association for Information Systems, 56, pp-pp. Retrieved from https://aisel.aisnet.org/cais/vol56/iss1/7

Larsen, B., & Narayan, J. (2023, January). *Generative AI – a game-changer society needs to be ready for | World Economic Forum*. https://www.weforum.org/stories/2023/01/davos23- generative-ai-a-game-changer-industries-and-society-code-developers/

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21(2).

Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022). News from Generative Artificial Intelligence Is Believed Less. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 97–106.

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digital Medicine*, 6(1), 120.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), 1–41. DOI: 10.1145/3425780

Mitra, A., Mohanty, S. P., Corcoran, P., & Kougianos, E. (2021). A machine learning-based approach for deepfake detection in social media through key video frame extraction. *SN Computer Science*, 2(2), 98. DOI: 10.1007/s42979- 021- 00495- x

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. *AI and Ethics*.

Muhly, F., Chizzonic, E., & Leo, P. (2025). AI-deepfake scams and the importance of a holistic communication security strategy. *International Cybersecurity Law Review*, *6*, 53–61. https://doi.org/10.1365/s43439-025-00143-7

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief History of Artificial Intelligence. *Neuroimaging Clinics of North America*, 30(4), 393–399.

Pavlik, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, 78(1), 84–93.

Robinson, O. C. (2013). Sampling in Interview-Based Qualitative Research: A Theoretical and Practical Guide. *Qualitative Research in Psychology*, *11*(1), 25–41. https://doi.org/10.1080/14780887.2013.801543

Ruane, E., & Birhane, A. (2019). Conversational AI: Social and Ethical Considerations. *AICS - 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*.

SecureWorld. (2024, February 13). Hong Kong Clerk Defrauded of $25 Million in Sophisticated Deepfake Scam. Retrieved from https://www.secureworld.io/industry-news/hong-kong-deepfake-cybercrime

Sherman, J. (2024). A Feast of Fraud: How International Hesitations to Regulate Deepfakes are Creating a Buffet for Financial Crimes. *The George Washington International Law Review*, *56*(1 & 2), 91–118. https://perma.cc/D3ZY-SJ4U]

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, 614(7947), 214–216.

Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., Liu, Z., Liu, Y., Wang, Y., Zhang, Z., Vidgen, B., Kailkhura, B., Xiong, C., Xiao, C., … Zhao, Y. (2024). TrustLLM: Trustworthiness in Large Language Models.

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*, *34*(2), 399–408. https://doi.org/10.1287/ISRE.2023.ED.V34.N2

The Straits Times. (2024, February 4). HK firm scammed of $34 million after employee duped by video call with deepfake of CFO. Retrieved from https://www.straitstimes.com/asia/east-asia/hk-firm-scammed-of-34-million-after-employee-is-duped-by-video-call-with-deepfake-of-cfo

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313–313.

Trend Micro (HK). (2024, February 7). Deepfake CFO Video Calls Result in $25MM in Damages. Retrieved from https://www.trendmicro.com/en_hk/research/24/b/deepfake-video-calls.html

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1). https://doi.org/10.1177/2056305120903408 (Original work published 2020)

Velasquez, M. G.: 2002, Business Ethics: Concepts and Cases (Prentice-Hall of India, New Delhi).

VOA. (2024, February 4). Deepfake Scam Video Cost Company $26 Million, Hong Kong Police Says. Retrieved from https://www.voanews.com/a/deepfake-scam-video-cost-company-26million-hong-kong-police-says/7470542.html

Wagner, T., & Blewer, A. (2019). "The word real is no longer real": Deepfakes, gender, and the challenges of AI-altered video. *Open Information Science*, 3(1), 32–46. DOI: 10.1515/opis- 2019- 0003

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity.

Zohny, H., Mcmillan, J., & King, M. (2023). Ethics of generative AI. *J Med Ethics*, *49*, 79. https://doi.org/10.1136/jme-2023-108909

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity.

Zohny, H., Mcmillan, J., & King, M. (2023). Ethics of generative AI. *J Med Ethics*, *49*, 79. https://doi.org/10.1136/jme-2023-108909

INFORMATION RESILIENCE & INTEGRITY SYMPOSIUM

Generative AI and Information Resilience in the Asia-Pacific: Actions and Adaptations

Faculty of Social and Political Sciences
Universitas Gadjah Mada

21 August 2025

saferinternetlab.org/iris · @saferinetlab · @cfds_ugm · iris@saferinternetlab.org