中國行政評論　第 32 卷第 2 期
The Chinese Public Administration Review
Vol.32 No.2 June 2026.36-61
DOI:10.6635/cpar.202606_32(2).0002

The Ethical Considerations between Artificial Intelligence and Public Policy in the Wake of ChatGPT: A Humanistic Approach[1]

Chunmeng Lu*

Abstract

This paper investigates the ethical implications of Generative Artificial Intelligence (GAI) in public policy through a humanistic lens. By comparing current AI advancements to historical scientific controversies, the study contextualizes the governance challenges addressed by proactive regulations like the EU AI Act (2024).
The research identifies critical ethical risks—including algorithmic bias, misinformation, privacy infringement, and transparency erosion—arguing that these issues stem from a systemic neglect of "humanity-first" principles. Through analysis of healthcare, finance, and autonomous systems, the study highlights specific threats such as diagnostic bias and safety concerns. To mitigate these risks, the paper advocates for a framework grounded in individual autonomy, fairness, and accountability.
It concludes that a multi-stakeholder approach involving policymakers, developers, and civil society is essential to safeguard human values while harnessing GAI's potential. Future research should focus on the long-term societal impacts of the evolving human-AI relationship.

Keywords: Generative AI, Public Policy, AI Ethics, Humanistic Approach, EU AI Act, Data Governance

[1] Paper presented at the 2025 ASPA Annual Conference, Washington, D. C., March 28-April 1, 2025.
*Associate Professor, Department of Public Management and Policy, Tung-Hai University. E-mail: lucg@thu.edu.tw Received: March 15, 2026. Accepted: May 18, 2026.

I. **Introduction**

The emergence of ChatGPT in late 2022 signaled a transformative era for Generative Artificial Intelligence (GAI), sparking both global enthusiasm and critical concern regarding its ethical and security implications. Beyond the initial technooptimism, figures such as Hinton (2024) have warned that GAI systems, through their mastery of human psychology and political machinations, could eventually engage in autonomous self-modification. This trajectory suggests a potential evolution from human-centric tools to autonomous entities with self-derived objectives, representing a fundamental challenge to existing ethical frameworks. This technological acceleration mirrors past ethical crises in science, most notably the 2018 genome-editing controversy involving Dr. He Jiankui (Alonso & Savulescu, 2021). The international backlash against that project highlighted the fragility of scientific integrity when individual ambition bypasses normative boundaries. Today, AI poses a similar dual-use threat, necessitating proactive governance such as the European Union's AI Act (2024). This legislation reflects a global imperative to internalize ethical considerations within the AI development lifecycle.

The capacity of GAI to perform complex cognitive tasks—once the sole domain of human intelligence—is the result of a long evolutionary process. From mid-20thcentury symbolic reasoning (Russell & Norvig, 2021) to the modern era of deep learning catalyzed by high-performance computing (HPC), AI has transitioned from rigid, rule-based systems to fluid, data-driven architectures. This study seeks to analyze the ethical dimensions of this transition, providing an evidentiary framework for policies that foster technological innovation while safeguarding against societal and existential risks.

These ethical risks find a modern parallel in the accelerated evolution of Artificial Intelligence (AI), which presents analogous systemic dangers if left unregulated. In a proactive response to these emerging sociotechnical threats, the European Union ratified the AI Act (2024). This legislation establishes a comprehensive regulatory framework intended to internalize ethical considerations and mitigate the deleterious impacts of future AI trajectories.

Central to this technological shift is Generative Artificial Intelligence (GAI), which denotes the development of computational architectures capable of executing high-level cognitive functions—such as autonomous learning, multi-dimensional problem-solving, and strategic decision-making—that were traditionally considered the exclusive domain of humanistic inquiry. While contemporary in its impact, the conceptual foundations of GAI are rooted in mid-20th-century advancements in

symbolic reasoning, expert systems, and heuristic algorithms (Russell & Norvig, 2021). Facilitated by the exponential expansion of the Internet and the integration of large-scale data platforms, GAI has undergone a significant paradigm shift: transitioning from rule-based expert systems and rudimentary data mining to the current dominance of machine learning and deep learning. This evolutionary progress has been catalyzed primarily by breakthroughs in neural architectural design and the proliferation of high-performance computing (HPC) resources.

This study employs a Qualitative Systematic Review and Policy Discourse Analysis to evaluate the efficacy of current mitigation strategies for Generative AI (GAI). The methodology is structured into three primary phases: A. Literature Synthesis and Taxonomization A comprehensive review of peer-reviewed literature (2016–2024) from computer science, ethics, and jurisprudence was conducted. This facilitated the categorization of mitigation efforts into technical, organizational, and normative dimensions. B. Comparative Regulatory Analysis The study analyzes key global policy instruments, such as the EU AI Act, the UNESCO Recommendation on the Ethics of AI, and the Asilomar AI Principles, to identify convergence and divergence in international governance standards. C. Socio-Technical Framework Evaluation Using a "Responsible Research and Innovation" (RRI) lens, the study assesses how technical solutions (e.g., LIME, debiasing algorithms) intersect with human oversight and socio-economic safety nets to address the structural risks identified in the preceding sections.

Ultimately, this study seeks to synthesize a comprehensive understanding of GAI-related ethics, providing an evidentiary basis for the formulation of responsible AI policies that balance technological innovation with the mitigation of existential and societal risks.

II. Literature Review

GAI represents a paradigmatic shift in artificial intelligence capabilities, specifically through algorithms designed to synthesize novel content—ranging from textual and visual media to complex programming code—by leveraging vast datasets accessible via high-speed network infrastructures. The defining characteristics of GAI can be categorized as follows:

Firstly, the GAI systems demonstrate a capacity for producing novel outputs that simulate human-like creativity (Goodfellow et al., 2020). This is called the generative ingenuity. However, this "creativity" remains functionally contingent upon the scope

of accessible training data. Current architectures lack autonomous cognition; thus, their outputs are emergent properties of data synthesis rather than manifestations of independent thought. The quality of the dataset utilized by the GAI system determines the value of the output.

Secondly**, r**ather than merely replicating existing data, GAI facilitates the generation of unique configurations by identifying and reorganizing latent patterns within information. To an unrefined observer, these outputs of unique structural innovation appear entirely original; however, they are more accurately described as sophisticated recontextualizations of pre-existing datasets.

Finally, the GAI models are primarily trained on massive, heterogeneous datasets, enabling the extraction of complex correlations and underlying structures (LeCun et al., 2015). While this data-driven nature of data centricity constitutes the core strength of GAI, it simultaneously represents a fundamental limitation, as the parameters of the training corpus inherently bind the quality and bias of the output.

The technological trajectory of GAI has experienced an accelerated surge, underpinned by advancements in deep learning and the proliferation of sophisticated LLMs. As these systems achieve remarkable proficiency, their influence is transforming diverse sectors—ranging from the precision of healthcare to the strategic complexities of modern warfare. The latter, in particular, underscores the urgency of evaluating the intersection of autonomous technology and human ethics. Consequently, this study pivots on a fundamental inquiry: In what ways has the development of GAI necessitated a re-evaluation of ethical paradigms, and what are its long-term implications for global humanistic values?

The rapid evolution of the GAI introduces a unique set of ethical dilemmas that challenge established paradigms of accountability, distributive justice, and the ontological nature of human creativity. This research investigates the specific modalities through which GAI development impacts ethical considerations across diverse domains. Addressing these challenges is imperative for the following reasons:

The most serious challenge to humanity is that GAI technologies exert a profound influence on personal spheres, ranging from the transformation of labor markets and employment opportunities to the erosion of data privacy and psychological well-being. Identifying and mitigating these ethical risks is essential to ensure that technological deployment remains anthropocentric and to safeguard individual rights and interests.

The second threat of the GAI is the societal and structural impact on personal daily lives. The pervasive adoption of GAI has the potential to reshape socio-economic frameworks and the foundational fabric of collective life. A robust ethical response is necessary to ensure that AI-driven innovation promotes social equity and averts systemic deleterious consequences. Yet the results of current progress associated with GAI tend to care more about a business's fundamental interests, rather than individual rights and livelihood.

Ideally, the ethical governance of GAI is fundamental to fostering public trust and institutional legitimacy. Yet the progress made by the technological world tends to focus more on technological growth, while ignoring sustainable development in the GAI fields. By proactively addressing ethical concerns, the scientific and policy communities can preempt potential crises, thereby ensuring the long-term sustainability and responsible advancement of autonomous systems.

Consequently, this research investigates the ethical dimensions of GAI, and its multifaceted applications through a human-centric lens. The specific objectives of this study are as follows:

The paper will firstly go through a systematic ethical analysis to critically analyze the ethical implications of GAI development and deployment across key sectors, including healthcare, finance, education, and autonomous systems.

Then it will identify related ethical dilemmas based on the principle of humanity, and investigate discrete ethical challenges precipitated by GAI with a specific focus on algorithmic bias, socio-economic discrimination, the proliferation of misinformation, structural job displacement, privacy infringements, and the "black box" nature of system opacity.

Formulation of Mitigation Strategies: To explore robust mitigation frameworks and best practices, including the institutionalization of ethical guidelines and the promotion of responsible research protocols. This objective encompasses an evaluation of mandatory data integrity standards for internet-scale datasets and the role of state-level regulatory interventions in preempting malevolent exploitation by adversarial actors.

Evaluation of Regulatory Frameworks: To examine extant ethical and legal frameworks governing AI, assessing their efficacy and adaptability in addressing the unique socio-technical challenges inherent in GAI.

Identification of Research Lacunae: To identify existing knowledge gaps and propose trajectories for future inquiry regarding the intersection of GAI with social, ethical, and jurisprudence-based implications.

A. Benefits and Challenges of GAI

The rapid evolution of GAI has catalyzed a broad academic and industrial consensus regarding its transformative potential. When developed and deployed within a responsible framework, GAI offers substantial advancements across multiple dimensions of human society. The prevailing literature identifies several key areas of positive impact:

The most significant progress occurs in the healthcare business with a paradigm shift. GAI is poised to revolutionize the biomedical sector by facilitating personalized precision medicine, accelerating in silico drug discovery, and enhancing diagnostic accuracy and treatment (Philips, 2024; Rivas, 2019). Beyond clinical applications, medical professionals leverage the high-velocity computational and predictive analytics of GAI to optimize complex treatment protocols and patient management systems (Philips, 2024; Rivas, 2019).

Another great feature of GAI is that it is catalyzing economic growth and resilience for the world. From a macroeconomic perspective, GAI drives economic growth and expansion through the automation of cognitive tasks, significant productivity gains, and the emergence of novel industries (McKinsey, 2023; Nikravesh, 2025). Furthermore, GAI's capacity for predictive modeling equips corporations with the analytical tools necessary to navigate market volatility and systemic economic uncertainty, thereby fostering institutional resilience (McKinsey, 2023; Nikravesh, 2025).

GAI can also advance environmental sustainability by formulating more efficient designs of the energy consumption systems. GAI contributes to ecological preservation by optimizing energy consumption patterns and refining resource management systems. In the realm of public policy, GAI enables the formulation of innovative energy strategies by simulating the optimal allocation of diverse energy sources, thus providing a data-driven foundation for sustainable development (Enel, 2024; ResearchGate, 2025).

GAI can also enhance human capital and quality of life for the general public. GAI enhances individual and communal well-being by democratizing access to tailored educational resources and improving accessibility for marginalized populations through

universal access to the internet. By enabling hyper-personalized user experiences, these technologies facilitate a more inclusive digital environment (ZOOZ Consulting, 2024; Pinky Nandwana, 2025).

When it comes to making better public policy and governance, GAI helps policymakers tackle complex social issues by combining diverse data points into practical, coherent policy options. By including multiple stakeholder views and testing different socio-economic scenarios, GAI acts as a vital tool for evidence-based governance and social problem-solving (Sabherwal & Grover, 2024; OECD, 2022).

B. Ethical Challenges in the Advancement of GAI

Even though there are so many benefits associated with GAI, the discursive landscape regarding GAI's progress is characterized by a critical examination of its ethical dilemmas. Prevailing scholarship identifies several systemic risks that could impede human progress if left unaddressed.

The first risk is algorithmic bias and systematic discrimination against people. GAI systems, trained on heterogeneous datasets, often inadvertently internalize and amplify extant societal prejudices concerning gender, race, and socio-economic status. A seminal study by Buolamwini and Gebru (2018) revealed that facial recognition technologies exhibited significant racial and gender bias; for instance, error rates were as high as 34.7% for darker-skinned females, compared to a maximum error rate of 0.8% for lighter-skinned males. Beyond mere reflection, GAI can "normalize" these biases by generating synthetic content that reinforces harmful stereotypes (Crawford, 2016).

The second danger is information integrity revealed by misinformation and deepfakes. The capacity for GAI to produce hyper-realistic synthetic media—commonly termed "deepfakes"—poses a fundamental threat to epistemic trust. This technology facilitates the dissemination of sophisticated propaganda and manipulated media, which can undermine democratic processes and exacerbate social polarization (Howard et al., 2020). The low barrier to entry for creating fabricated news articles and audio-visual content complicates the distinction between empirical reality and algorithmic construct (Deepfakes, 2023).

Thirdly, structural socio-economic disruption caused by employment and job market. The automation potential of GAI raises significant concerns regarding job displacement across cognitive-heavy sectors such as data analysis and content creation. While technology may catalyze new labor markets, the velocity of this transition may

outpace workforce readaptation, potentially leading to increased income inequality and social instability (Acemoglu & Restrepo, 2018).

Privacy infringement and Surveillance capitalism are the fourth risk as technological progress endangers the general public's daily lives. The GAI development cycle frequently relies on the extraction of vast personal datasets, often without explicit informed consent. This raises critical concerns regarding unintended surveillance and the erosion of civil liberties. As noted by Zuboff (2019), the enhancement of surveillance capabilities through AI enables state and corporate entities to monitor human behavior on an unprecedented scale, necessitating a robust framework for data protection (Solove, 2008).

Because of the inherent deficit of GAI and the dataset, opacity and the phenomenon of hallucination are always threatening the utility of applications of GAI. Many deep learning models function as "black boxes," lacking interpretability and explainability (Lipton, 2018). A critical ethical concern is the tendency for GAI to "hallucinate"—generating factually incorrect but linguistically confident responses to avoid output gaps. This "face-saving" mechanism (hallucination) compromises the credibility of the information ecosystem, misleading users and necessitating rigorous verification protocols.

Human autonomy and moral agency pose threats to humanity in general. As GAI systems assume greater autonomy in critical sectors like healthcare and autonomous transport, they challenge traditional notions of moral agency and liability (Floridi, 2016). Ensuring human-in-the-loop (HITL) oversight remains essential to maintaining individual control and allocating responsibility in the event of system malfunction.

The primary objective of this inquiry is to substantiate the ethical discourse surrounding GAI and to offer a scaffold for evidence-based policy development. By bridging the gap between rapid technological evolution and normative oversight, this study advocates for an integrated approach to responsible AI. The goal is to facilitate a synergy between innovation and ethical safeguards, ensuring that the deployment of these technologies aligns with broader societal interests.

C. Human Ethical Frameworks for the Governance of Generative Artificial Intelligence

Building upon the previously articulated risks associated with Generative Artificial Intelligence (GAI), this section elucidates the core ethical pillars based on humanity requisite for the responsible development and deployment of these systems.

The first principle deals with individual autonomy and agency. This principle prioritizes the preservation of human self-determination amidst the proliferation of sophisticated AI architectures. It posits that GAI should function as an augmentative tool rather than a manipulative force (Floridi, 2016; Sparrow, 2016). Critical dimensions include: (1) mitigation of undue influence: systems must be engineered to preclude coercive behaviors, such as algorithmic persuasion or the construction of "filter bubbles" that restrict cognitive diversity (Sunstein, 2020); (2) maintenance of Human-in-the-Loop (HITL) oversight. Research priorities should consistently emphasize that humans must retain ultimate jurisdictional control over decision-making processes, particularly in high-stakes autonomous systems (Wallach & Allen, 2009). Humans should be granted the final power of shutting down the system--the ultimate final resolution; (3) promotion of human flourishing shall remain the ultimate goal for any GAI progress. The deployment of GAI should be oriented toward enhancing human well-being and expanding human capabilities rather than displacing agency (Floridi, 2016).

The second guideline for GAI progress shall emphasize fairness and non-discrimination. The principle of equity mandates that GAI systems be developed through a lens of social justice, ensuring they do not perpetuate systemic inequalities. (Friedman & Nissenbaum, 1996). Further development of GAI shall engaging in activities involving the following: (1) algorithmic bias mitigation: It is imperative to identify and neutralize biases inherent in training datasets and model architectures that may lead to disparate impacts (Buolamwini & Gebru, 2018). (2) equitable access: Efforts must ensure that the benefits of GAI are distributed across diverse sociodemographic strata, preventing a "digital divide" based on socioeconomic status (Floridi & Cowls, 2019). (3) prevention of discriminatory outcomes: stringent safeguards are required in sensitive domains such as judicial sentencing, credit lending, and recruitment to prevent automated discrimination (Selbst & Selbst, 2019).

The third principle is to seek adequate transparency and explainability of GAI models. Transparency serves as a foundational requirement for establishing public trust and systemic accountability (Goodman & Flaxman, 2017). It requires the model interpretability that demands the current research to focus on transitioning from "black box" architectures to interpretable models that allow for human insight into the logic of AI-generated outputs (Lipton, 2018). Furthermore, it asks the provision of meaningful

explanations for suggestions made by GAI models that require clear, concise rationales for GAI decisions to facilitate informed interaction and contestability for stakeholders (Adadi & Berrada, 2018). It also stresses the need for open research Paradigms to foster a transparent development environment through collaborative research, which is essential for long-term accountability (Mittelstadt et al., 2016).

D. Privacy and Data Protection for Humanity

Given that GAI relies on ingesting large datasets, strong privacy measures are crucial (Solove, 2008). However, to prevent dataset misuse, data minimization is necessary. Developers should follow the principle of parsimony, collecting only the data points needed for effectiveness. Cybersecurity is the top concern for protecting privacy. Building cybersecurity resilience involves using advanced encryption and security protocols to protect personal data from unauthorized access. Any program created under GAI guidelines must include privacy-preserving architectures. Incorporating techniques like differential privacy and federated learning enables model training without compromising individual anonymity (Shokri & Shmatikov, 2015).

E. Responsibility and Accountability of Agency

A comprehensive ethical framework necessitates clear attribution of responsibility across the AI lifecycle (Floridi, 2016). It shall demand that actors involved in all aspects of GAI development follow the following regulations. Mechanism to unsure identification of responsible actors shall be mandatory. Jurisprudential and ethical clarity is required to delineate the duties of developers, deployers, and end-users if legal issues are involved. There are also accountability mechanisms to be established by formal and legal protocols to address harm and ensure that responsible parties remain answerable for the societal impacts of their systems. The identification of principles of humanity-first shall be instilled not only in education systems but also in all aspects of all institutions. The concepts of lifecycle ethics promote a culture of "Ethics by Design" that ensures that moral considerations are integrated from the initial conceptualization through to the decommissioning of GAI systems (Floridi & Cowls, 2019).

In summary, the rapid evolution of Generative AI represents a dual-edged sword for modern society, offering transformative breakthroughs in healthcare, economic resilience, and environmental sustainability while simultaneously introducing systemic ethical risks. While GAI's capacity for predictive modeling and cognitive automation can catalyze a paradigm shift in human productivity and quality of life, these advancements are shadowed by the persistence of algorithmic bias, the erosion of

information integrity through deepfakes, and the potential displacement of human agency. The documented error rates in facial recognition and the emergence of "black box" hallucinations underscore that technical proficiency does not inherently equate to social reliability. Therefore, the trajectory of GAI must be navigated with a profound understanding that technological utility is secondary to the preservation of fundamental human rights and epistemic trust.

To ensure that GAI serves as a catalyst for human flourishing rather than a source of structural inequality, it is imperative to implement a robust ethical framework rooted in transparency, accountability, and the "human-in-the-loop" principle. By prioritizing "Ethics by Design," developers and policymakers can mitigate privacy infringements and systemic discrimination before they become entrenched in social infrastructure. The ultimate goal of GAI governance should not be to stifle innovation, but to foster a synergistic relationship where automated intelligence augments human capability within a secure, equitable, and interpretable environment. Moving forward, the global community must maintain rigorous oversight and inclusive discourse to ensure that the deployment of these powerful technologies aligns consistently with the collective interests of humanity.

F. Sectoral Ethical Implications of GAI: High-Risk Domains

While the theoretical benefits of Generative Artificial Intelligence (GAI) are substantial, its implementation across critical business domains introduces significant ethical vulnerabilities that necessitate rigorous scholarly attention.

(A) Healthcare: Diagnostic Bias and Data Vulnerability: The integration of GAI into clinical environments—ranging from medical pedagogy to diagnostic oversight—presents acute ethical challenges. The first challenge is algorithmic bias in diagnostics. GAI-driven diagnostic tools risk inheriting and amplifying systemic biases present in historical training datasets. For instance, models trained predominantly on Caucasian cohorts frequently demonstrate diminished accuracy when diagnosing conditions in non-white populations. Research indicates that such algorithmic disparities can lead to a 20% to 50% reduction in diagnostic precision for marginalized groups, thereby exacerbating existing health inequities (Obermeyer et al., 2019). The second challenge comes from privacy in precision medicine. The shift toward personalized medicine requires the ingestion of granular patient metrics, including genomic data and longitudinal lifestyle records. This high-density data collection elevates the risk of catastrophic data breaches. Unauthorized access or the secondary misuse of sensitive

information poses a fundamental threat to patient confidentiality and the foundational trust of the clinician-patient relationship (Egelman et al., 2017).

(B) Finance: Systemic Instability and Credit Equity: The financial sector's transition to ICT-driven environments has accelerated the adoption of GAI, introducing new forms of market and social risk. An automatic trading mechanism may trigger high-frequency trading and market volatility. GAI-powered algorithms can intensify market fluctuations. The "black box" nature of these high-frequency models obscures market behavior, heightening the probability of systemic failures and "flash crashes" (Kirilenko et al., 2017). The said mechanism might also trigger algorithmic discrimination in credit scoring for minority people. Automated credit assessments often perpetuate socioeconomic disparities. By relying on proxy variables that correlate with protected characteristics, these models may unfairly penalize individuals from lower-income brackets or underserved geographic regions. Studies on mortgage algorithms have shown that minority applicants can be 40% to 80% more likely to be denied conventional loans compared to white applicants with similar financial profiles (Barocas & Selbst, 2016). The same institute may also cause systemic financial risk with the widespread deployment of interconnected GAI systems, which increases the complexity of global finance, creating a "cascading failure" risk where a single algorithmic error could trigger a broader economic crisis.

(C) Autonomous Systems: Safety, Liability, and Surveillance: As automation permeates manufacturing and transport, the ethical "alignment problem" becomes a matter of physical safety. The first ethical dilemma deals with accidents caused by autonomous vehicles (AVs). AVs face complex moral imperatives, such as prioritizing whose lives in unavoidable collision scenarios (The "Trolley Problem" in AI). Determining liability and establishing a clear chain of accountability in the event of an accident remains a significant jurisprudential hurdle when human lives are involved (Lin, 2013). Surveillance and lethal autonomy represent another dilemma when the use of drones and autonomous mechanisms raises profound concerns regarding pervasive surveillance and the erosion of privacy. Furthermore, the development of Lethal Autonomous Weapon Systems (LAWS) introduces the risk of "dehumanized warfare," where life-and-death decisions are delegated to non-human actors (Sharkey, 2012). Labor displacement needs to be addressed in normal times, when in sectors such as logistics and transportation, the rapid transition to autonomous systems threatens significant job displacement, necessitating a proactive policy response to manage the transition of the human workforce.

(D) Synthesis-The Ultimate Mandate for Human Oversight: To mitigate the

potential for GAI-related mechanisms to endanger human welfare, a framework prioritizing safety, accountability, and human oversight is essential. This requires rigorous empirical testing to ensure the reliability of systems through continuous monitoring and fail-safe mechanisms; legal attribution of responsibility to establish clear lines of accountability for developers and operators to address harm (Wallach & Allen, 2009). And issues involve Human-in-the-Loop (HITL) integration to maintain a "human-centric" design philosophy, ensuring that technological operations remain aligned with fundamental human values and ethical standards.

In summary, the integration of Generative AI (GAI) into healthcare and finance reveals deep-seated risks regarding algorithmic bias and data security. In healthcare, models trained on narrow datasets show a 20% to 50% reduction in diagnostic precision for non-white populations, while the collection of genomic data for precision medicine increases the risk of catastrophic privacy breaches. Similarly, in the financial sector, high-frequency "black box" algorithms can trigger systemic market instability. These tools also perpetuate socioeconomic gaps; for instance, studies indicate that minority applicants are 40% to 80% more likely to be denied conventional loans than white applicants with equivalent financial profiles.

Beyond data and equity, the rise of autonomous systems introduces critical physical and legal dilemmas. Autonomous vehicles (AVs) face the "Trolley Problem," in which moral imperatives in unavoidable collisions pose complex liability hurdles. The shift toward automation also threatens the workforce in sectors like logistics, necessitating proactive policies to manage labor displacement. More severely, the development of Lethal Autonomous Weapon Systems (LAWS) and pervasive drone surveillance raises the specter of "dehumanized warfare," where life-and-death decisions are removed from human hands, and privacy is fundamentally eroded.

To mitigate these ethical vulnerabilities, the text advocates for a comprehensive mandate for human oversight and legal accountability. This framework emphasizes rigorous empirical testing and the implementation of fail-safe mechanisms to ensure system reliability. By establishing clear lines of legal responsibility for developers and adopting a Human-in-the-Loop (HITL) design philosophy, organizations can ensure that GAI operations remain aligned with human values. Ultimately, the goal is to balance technological advancement with the protection of human welfare through continuous monitoring and ethical alignment.

III. Strategic Mitigation Frameworks and Public Policy

The multi-faceted nature of the threats posed by Generative Artificial Intelligence (GAI) necessitates a collaborative, cross-sectoral response. Stakeholders—including software engineers, corporate entities, end-users, and regulatory bodies—must synchronize efforts across the development, deployment, and oversight phases. The following analysis delineates the primary mitigation strategies essential for ethical governance.

A. Data Governance and Algorithmic Equity

Effective mitigation begins with the structural integrity of the data lifecycle. It begins with pre-processing of data with privacy preserving as the priority. Implementing techniques to anonymize or mask personally identifiable information (PII) is a fundamental prerequisite. However, as de-identification often remains incomplete, practitioners must remain vigilant against residual re-identification risks (Narayanan & Shmatikov, 2008). Secondly, stakeholders must exercise with care in implementing bias detection and algorithmic recourse, as it is imperative to deploy advanced debiasing algorithms utilizing fairness constraints and counterfactual modeling. These tools allow for "algorithmic recourse," ensuring that model outputs do not disproportionately disadvantage specific cohorts (Hardt et al., 2016; Kusner et al., 2017). Thirdly, workforce diversification must be carried out to preemptively identify "blind spots" in model design. Development teams must reflect sociodemographic diversity. Heterogeneous teams are empirically more likely to recognize and mitigate cultural and systemic biases that remain invisible to homogeneous groups (Crawford, 2016).

B. Structural Transparency and Explainability

To foster institutional and public trust, GAI systems must transition toward "glass-box" architectures. Developers should prioritize feature importance analysis and local interpretable model-agnostic explanations (LIME). These methods extract human-readable logic from complex neural networks, allowing users to understand the "why" behind an AI-generated prediction (Ribeiro et al., 2016). Integrating human-centric oversight into the deployment phase ensures that model performance remains aligned with ethical expectations and societal norms (Amershi et al., 2019).

C. Normative Development and Global Standards

While the competitive drive for technological primacy is significant, it must be moderated by adherence to international ethical benchmarks. Organizations should adopt established frameworks, such as the *Asilomar AI Principles* and the *UNESCO*

Recommendation on the Ethics of Artificial Intelligence, to guide their R&D trajectories (Future of Life Institute, 2017; UNESCO, 2021). Pre-deployment risk assessment must be thoroughly observed before robust testing protocols, which include independent ethical audits and safety evaluations to ensure models meet reliability standards before market entry.

D. The Crucial Role of Regulatory Governance

Policymakers and non-governmental organizations (NGOs) should serve as the final arbiters of ethical AI utility. Their mandate includes a well-established legislative framework with national and international regulations (e.g., the EU AI Act) to ensure the establishment of clear standards for accountability and transparency (Goodman & Flaxman, 2017). And social safety nets need to be created to address the socioeconomic disruptions caused by GAI, specifically through workforce retraining programs and policies aimed at mitigating income inequality resulting from automation. Most importantly, a comprehensive human-rights-based protection mechanism needs to be institutionalized to ensure that AI deployment does not infringe upon fundamental human rights or individual privacy, maintaining a balance between technological innovation and civil liberties.

IV. Strategic Policy Options for Regulators and Governing Bodies

To effectively govern the rapid evolution of Generative Artificial Intelligence (GAI), policymakers must transition from passive observation to active, multi-layered regulation. The following framework outlines specific policy instruments designed to institutionalize ethical oversight.

An utmost urgent task would be to create adequate statutory regulatory frameworks at all governmental levels. Governments and international oversight boards must prioritize codifying AI governance into enforceable law. This includes: (A) Algorithmic Accountability: Enacting mandates that require developers to provide "right to explanation" for AI-driven decisions, particularly in high-stakes sectors like healthcare and autonomous systems (Goodman & Flaxman, 2017). (B)Sector-Specific Oversight: Tailoring data protection and safety regulations to the unique risk profiles of critical industries, ensuring that GAI deployment does not bypass existing professional standards.

Establishing industry-wide standards is a critical priority for both state actors and non-governmental organizations (NGOs). The first challenge is to define standardized protocols for data quality and bias mitigation to ensure consistency in safety audits

across the GAI landscape. Construct a collaborative governance to promote a "public-private-NGO" partnership model where civil society acts as a monitor, helping state bodies refine laws that keep pace with technological iterations.

The establishment of formal Ethical Review Boards (ERBs) is necessary to evaluate the societal impact of AI projects before market entry both locally and globally. On an international scale, ERBs and NGOs must exercise rigorous oversight over the militarization of AI. Given the existential risk posed by autonomous weapon systems, global monitoring of "AI-related weapons of mass destruction" is paramount to preventing dehumanized warfare. Public investment should be strategically redirected toward "AI for Social Good" while actively discouraging malevolent applications. Prioritizing funding for GAI applications in environmental sustainability, public health, and education to ensure technological dividends are shared equitably. Policymakers should implement strict bans or heavy surveillance on the development of lethal autonomous mechanisms, treating the weaponization of GAI as a significant threat to global stability.

Continuous investment in human capital and whistleblower protection to prevent misuse of GAI. Recognizing that humans should ultimately manage technology, policy must address the workforce's role in the AI lifecycle. Investing in comprehensive education programs to equip the workforce with the cognitive and technical skills required in an AI-integrated economy. Enacting robust legal protections for employees who identify and report ethical breaches or "algorithmic wrongdoings" to the authorities. This turns the workforce into a distributed layer of ethical monitoring.

The emergence of Generative Artificial Intelligence (GAI) signifies a paradigmatic shift in computational capabilities, moving beyond traditional analytical frameworks toward the autonomous synthesis of novel content. This evolution is underpinned by sophisticated algorithms designed to generate text, visual media, and programming code by leveraging vast datasets via high-speed network infrastructures. At its core, GAI is characterized by what may be termed "generative ingenuity"—a capacity to produce outputs that simulate human creativity (Goodfellow et al., 2020). However, scholarly critique suggests that this creativity is fundamentally "functionally contingent." Lacking autonomous cognition, GAI systems function as engines of data synthesis rather than independent thought; their outputs are emergent properties of the underlying training architecture rather than manifestations of sentient innovation. Consequently, the quality and boundaries of the utilized datasets dictate the inherent value and limitations of the system's output.

Rather than mere replication, GAI facilitates "structural innovation" by identifying

and reorganizing latent patterns within information. To a superficial observer, these configurations may appear entirely original, yet they are more accurately described as sophisticated recontextualizations of pre-existing data. As LeCun et al. (2015) emphasize, while the data-centric nature of GAI enables the extraction of complex correlations, it also binds the system to the biases and parameters of its training corpus. This technological trajectory has reached a critical juncture where the proficiency of Large Language Models (LLMs) is transforming sectors ranging from healthcare precision to the strategic complexities of modern warfare. This rapid expansion underscores an urgent need to evaluate the intersection of autonomous technology and human ethics, necessitating a fundamental re-evaluation of established moral paradigms.

Current literature highlights a profound tension between technological advancement and humanistic values. On a personal level, GAI exerts influence over labor markets, data privacy, and psychological well-being, raising questions about how to maintain an anthropocentric deployment of technology. On a societal level, the pervasive adoption of these systems threatens to reshape socio-economic frameworks. A critical observation in contemporary discourse is the misalignment between commercial interests and individual rights; the technological world tends to prioritize rapid growth and market dominance over sustainable development and the safeguarding of livelihoods. While ethical governance is theoretically recognized as fundamental to fostering public trust, the proactive institutionalization of such guidelines remains secondary to the pursuit of technical proficiency. Despite the documented benefits—such as the revolutionizing of personalized medicine (Philips, 2024; Rivas, 2019), the stimulation of macroeconomic growth (McKinsey, 2023), and the optimization of energy systems for environmental sustainability (Enel, 2024)—a significant research gap remains. Existing studies predominantly focus on the functional potential or isolated risks of GAI, yet there is a lack of integrated analysis regarding the "socio-technical adaptability" of current ethical and legal frameworks. Specifically, there is a dearth of research on how to transition from abstract ethical principles to mandatory, state-level regulatory interventions that can effectively preempt malevolent exploitation and systemic discrimination in a "black box" environment. This study, therefore, seeks to bridge this gap by investigating the multifaceted ethical dimensions of GAI through a human-centric lens, aiming to formulate robust mitigation strategies that address the opacity and rapid evolution of these autonomous systems.

V. Findings and Discussion

This analysis outlines three core findings regarding Generative Artificial Intelligence (GAI) risk mitigation: the necessity of holistic data governance to address algorithmic bias (Narayanan & Shmatikov, 2008; Hardt et al., 2016; Crawford, 2016), the shift toward structural transparency through "glass-box" models (Ribeiro et al., 2016; Goodman & Flaxman, 2017), and the urgent requirement for global normative standards to address existential risks (Future of Life Institute, 2017; UNESCO, 2021). Effective AI governance requires blending technical fairness, such as counterfactual modeling and LIME, with human-centric oversight and international ethical frameworks (Kusner et al., 2017; Amershi et al., 2019).

A. Holistic Data Governance and the Mitigation of Algorithmic Bias

The primary finding indicates that ethical AI governance is predicated on a rigorous, multi-staged data lifecycle management strategy. This involves the integration of privacy-preserving pre-processing, the application of advanced debiasing mathematical constraints, and the intentional diversification of development teams. The results suggest that technical fixes alone—such as PII masking—are insufficient due to the persistent nature of re-identification risks. Furthermore, "algorithmic recourse" emerges as a critical mechanism for ensuring equity, allowing for the correction of model outputs that unfairly disadvantage specific sociodemographic cohorts. The necessity of protecting individual privacy through anonymization is a wellengineered prerequisite; however, as noted by Narayanan and Shmatikov (2008, pp. 20-25), the complexity of modern datasets often allows for residual re-identification, rendering traditional masking techniques partially ineffective. Consequently, the discussion must shift from simple data "cleaning" to a more robust framework of "algorithmic equity." Hardt et al. (2016, pp. 3315-3323) argue that fairness in machine learning requires supervised learning models to satisfy "equalized odds," ensuring that prediction errors do not cluster around protected groups. When coupled with counterfactual modeling (Kusner et al., 2017, pp. 587-597), developers can simulate alternative scenarios to determine if a model's decision would change if a protected attribute (like race or gender) were different. This technical rigor, however, must be balanced by human diversity. Crawford (2016, pp. 11-14) emphasizes that "blind spots" in AI are often a direct reflection of the homogeneity of the engineering workforce. Therefore, the "results" of a GAI system are not merely products of code but are socio-technical outputs that require heterogeneous teams to identify systemic biases that are invisible to a monocultural developer base. Effective mitigation, then, is a hybrid of mathematical fairness and sociodemographic representation.

B. Structural Transparency and the Institutionalization of "Right to Explanation"

Research into GAI deployment underscores a shift from "black-box" models to "glass-box" architectures. The findings suggest that for AI to be trusted in high-stakes sectors like healthcare or law, it must provide "local interpretable model-agnostic explanations" (LIME). This transparency allows human overseers to verify the logic behind specific predictions. Moreover, the results highlight that transparency is not just a technical feature but a regulatory requirement, often referred to as the "right to explanation," which bridges the gap between complex neural network operations and human-readable logic. The demand for explainability addresses the fundamental "opacity" of deep learning. Ribeiro et al. (2016, pp. 1135-1144) propose that by identifying which features are most important to a specific prediction, LIME allows users to trust the model's reasoning. This is crucial because a model might reach a "correct" conclusion for the "wrong" reasons (e.g., detecting a disease based on a watermark on an X-ray rather than the pathology). From a policy perspective, Goodman and Flaxman (2017, pp. 50-55) discuss how the EU's General Data Protection Regulation (GDPR) and similar frameworks like the EU AI Act effectively mandate this transparency. The "right to explanation" serves as a safeguard against "algorithmic wrongdoing," ensuring that individuals impacted by automated decisions can challenge the underlying logic. Furthermore, Amershi et al. (2019, pp. 1-13) argue that human-AI interaction must be grounded in "humancentric oversight," where the AI provides clear signals of its uncertainty, allowing humans to intervene when the model's performance deviates from ethical norms. This discussion reinforces the idea that transparency is the linchpin of accountability; without understanding the "why" behind an output, institutional and public trust cannot be sustained.

C. Global Normative Standards and the Regulation of Existential Risks

The analysis reveals that the competitive drive for "AI primacy" poses significant risks to global stability, particularly regarding the weaponization of AI and the erosion of human rights. The results advocate for a transition from passive observation to active, multi-layered regulation through international frameworks (UNESCO, Asilomar Principles) and formal Ethical Review Boards (ERBs). A critical finding is the urgent need for a "human-rights-based protection mechanism" to balance technological innovation with civil liberties, alongside strict bans on lethal autonomous weapon systems. The governance of GAI must transcend national borders to be effective. As outlined by the Future of Life Institute (2017, pp. 1-5), the Asilomar AI Principles

provide a normative baseline for long-term safety and shared prosperity. However, normative "soft law" is insufficient without the "hard law" of enforceable statutory frameworks. UNESCO (2021, pp. 12-18) further highlights that AI development should never occur in a vacuum; it must be guided by "independent ethical audits" before market entry. The discussion regarding the militarization of AI is particularly grave.

The potential for "dehumanized warfare" through autonomous mechanisms represents an existential threat that requires a global monitoring regime similar to those used for weapons of mass destruction. Beyond physical safety, the socioeconomic impact of GAI—such as income inequality and job displacement—must be addressed via social safety nets and workforce retraining. Ultimately, the role of policymakers is to ensure that AI serves the "social good" (e.g., environmental sustainability and public health) rather than malevolent interests. By protecting whistleblowers and institutionalizing ERBs, society can create a "distributed layer of ethical monitoring" that keeps pace with the rapid iterations of GAI technology.

VI. Conclusion

This study seeks to decode the intricate ethical landscape of Generative Artificial Intelligence (GAI) by examining the friction between rapid technological innovation and normative risk. It aims to identify systemic vulnerabilities—such as algorithmic bias, data privacy breaches, and the "black-box" nature of non-interpretable models—while analyzing how these risks manifest uniquely across critical sectors like healthcare and finance. The novelty of this research lies in its transition from passive ethical observation to an active, multi-layered governance framework. Unlike existing literature that focuses solely on technical fixes, this study proposes a hybrid collaborative model that synchronizes data-level interventions (debiasing) with enforceable legal codification. A key innovation is the integration of a "right to explanation" mandate for high-stakes autonomous systems, bridging the gap between technical transparency and legal accountability.

This paper contributes a strategic roadmap for both policymakers and the scientific community: 1. Policy & Regulation: It provides a blueprint for "Ethical Review Boards" and sector-specific oversight, advocating for international standards to prohibit malevolent applications, such as lethal autonomous weapons. 2. Social Impact: It shifts the discourse toward "AI for Social Good," emphasizing the protection of labor markets and the democratization of public discourse to ensure AI alignment with human values. 3. Future Academic Inquiry: By acknowledging the volatility of GAI architectures, the study establishes a preemptive foundation for longitudinal research into the long-term cognitive and societal effects of human-AI interaction.

This study has investigated the multifaceted ethical landscape of Generative Artificial Intelligence (GAI), delineating a complex interplay between technological innovation and normative risk. The primary findings are summarized as follows. Systemic ethical vulnerabilities are inevitable. They pose the greatest threat to the further development of GAI. GAI development introduces significant risks regarding algorithmic bias, the proliferation of misinformation, data privacy violations, and the pervasive "black-box" nature of non-interpretable models. Domain-specific fragility is common across all divisions. The ethical implications of GAI are non-uniform across sectors. Critical domains such as healthcare and finance face unique threats, ranging from diagnostic disparities to systemic financial instability and the erosion of accountability in autonomous systems. The necessity for integrated mitigation is the greatest challenge to all actors in the policy-making arena. Addressing these challenges requires a multi-layered strategy that synchronizes data-level interventions (debiasing) with structural transparency and responsible research paradigms. Regulatory actions are imperative for governments at the local and international levels. Policymakers and regulators serve as the essential architects of ethical AI, tasked with implementing enforceable guidelines and international standards to safeguard human interests. Broad societal transformation is necessary to handle the challenge posed by the advancement of GAI within all aspects of life. The impact of GAI extends beyond technical metrics, profoundly reshaping labor markets, educational structures, and the trajectory of social equity. The proposed framework advocates for a shift from passive AI observation to active, multi-layered regulation by codifying governance into enforceable law. Key measures include mandating algorithmic accountability—specifically the "right to explanation" in high-stakes sectors—and establishing sector-specific oversight to ensure AI aligns with existing professional standards. To maintain consistency, the framework suggests a collaborative model where governments, industry, and NGOs partner to set standardized protocols for data quality and bias mitigation.

Beyond technical standards, the strategy emphasizes global security and human-centric protections. It calls for the establishment of Ethical Review Boards to monitor societal impacts and strictly prohibit the development of lethal autonomous weapons. By redirecting public investment toward "AI for Social Good" and enacting robust whistleblower protections, the policy aims to empower the workforce as an ethical monitoring layer. This ensures that the benefits of GAI are distributed equitably while preventing its misuse in militarized or malevolent applications.

Despite the depth of this analysis, several constraints inherent to the current state of scientific inquiry must be acknowledged. Thematic scope remains a challenge to be conquered. While this paper addresses core ethical dilemmas, it cannot exhaustively

account for all emergent sociotechnical risks about the future. Technological volatility may evolve beyond the scope of the paper. The rapid, non-linear evolution of GAI means that new ethical challenges may materialize as the underlying architectures advance. Therefore, the paper can only serve as a preemptive discussion for future academic exploration. Sectoral concentration in certain business fields may be neglected. The primary focus on healthcare and finance may overlook unique ethical nuances present in other burgeoning fields of AI application.

To navigate the "unlimited scientific endeavors" of the AI era, future scholarly inquiry should prioritize the following trajectories. The exploration of holistic multi-stakeholder guidelines is required in the future. Research must focus on developing comprehensive, cross-sectoral regulatory frameworks. This includes fostering ethical self-awareness among developers, prioritizing societal welfare over corporate profit, and establishing a global moratorium on GAI-driven autonomous weaponry. Furthermore, nuanced ethical frameworks demand in-depth investigation to cope with future challenges. There is a critical need to design dynamic ethical architectures that can adapt to the evolving capabilities of large-scale generative models. Long-term societal and cognitive impacts necessitate supplementary investigation to reveal long-term effects. Longitudinal studies are required to assess the profound influence of GAI on employment, education, and the fundamental nature of human cognition and social interaction. Empirical bias mitigation demands an urgent response from governing entities to look for remedies immediately. Future work should refine robust, scalable methodologies for the automated detection and prevention of bias throughout the AI lifecycle. The process of democratization of public discourse is the keystone for stringent oversight of GAI progress. Fostering inclusive public engagement is essential to ensure that GAI development remains aligned with diverse human values and collective societal priorities.

References

Acemoglu, D., & Restrepo, P. (2018). *Artificial intelligence, automation, and work.* National Bureau of Economic Research.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Alonso, M., & Savulescu, J. (2021). He Jiankui´s gene-editing experiment and the non-identity problem. *Bioethics*, *35*(6), 563–573.

Amershi, S., Cakmak, M., Lee, S., Pfeffer, J., Douglass, B., See, A., Sarne,

A., & Horvitz, E. (2019). Guidelines for human-AI collaboration. *arXiv preprint arXiv:1901.08949*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–730.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Srivasta, P., Gopinath, S., Chen, Y., Chen, M., Khan, R., Toki, T., Kong, M., Judkins, D., Potter, S., Amodei, D., Clark, S., McCandlish, S., & Olah, C. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 81, 77–91.

CBS News. (2024, June 16). "Godfather of Artificial Intelligence" Geoffrey Hinton on the promise, risks of advanced AI. *CBS News*. cbsnews.com

Crawford, K. (2016). *The invisible hand of AI*. AI Now Institute.

Egelman, S., Khanna, K., Lin, P., & Bloomrosen, M. (2017). The ethics of artificial intelligence in healthcare. *Artificial Intelligence in Medicine*, *78*, 11–17.

Enel. (2024). *Artificial intelligence and the energy transition: How AI is driving sustainability*. Enel Group.

European Union. (2024). *Artificial Intelligence Act*. artificialintelligenceact.eu

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence*. europa.eu

Floridi, L. (2016). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Floridi, L., & Cowls, J. (2019). On the ethics of artificial intelligence for more people. *Mind*, *128*(511), 1081–1109.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *Communications of the ACM*, *39*(5), 24–31.

Future of Life Institute. (2017). *Asilomar AI principles*. futureoflife.org Goodfellow, I., Bengio, Y., & Courville, A. (2020). *Deep learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27,* 2672–2680.

Goodman, B., & Flaxman, S. (2017). European Union regulations on

algorithmic decision-making and a "right to explanation." *AI Now Institute*.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3323–3331.

Hinton, G. (2024, December 10). *Nobel Prize banquet speech*. Stockholm City Hall.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Howard, P. N., Kollanyi, B., Bradshaw, S., & Neudert, L.-M. (2020). *The computational propaganda project*. Oxford University Press.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirilenko, A. A., Kyle, A. S., Samadi, M., & Tütüncü, R. (2017). The impact of high-frequency trading on market quality. *Journal of Financial Markets*, *34*, 1–22.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, *30,* 4066–4076.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lin, P. (2013). The ethics of autonomous vehicles: The trolley problem revisited. *The Journal of Ethics*, *17*(1), 37–56.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57.

McKinsey & Company. (2023, June 14). *The economic potential of generative AI: The next productivity frontier*. McKinsey Global Institute. mckinsey.com

Mittelstadt, B., Taddeo, M., Floridi, L., & Welser, M. (2016). The ethics of artificial intelligence. *AI & Society*, *31*(2), 189–201.

Nandwana, P. (2025). AI for humanity: Enhancing quality of life through personalized experiences and accessibility. *International Journal of Innovative Research in Technology (IJIRT)*. *12*(1), 100-110.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of social networks. *Proceedings of the 2008 ACM Symposium on Principles of Database Systems*, 111–120.

Nikravesh, M. (2025). *Generative AI and the future of global economy: Productivity, automation, and new job markets*. Springer Nature.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in algorithmic risk scores. *Science*, *366*(6464), 447–453.

OECD. (2019). *OECD AI principles*. oecd.org

OECD. (2022, November). *AI and public policy: Improving government*

decision-making and service delivery. OECD Publishing.

Philips. (2024, June 18). *Future Health Index 2024 global report: Better care for more people*. Royal Philips.

ResearchGate. (2025). *Sustainable AI: Optimizing resource management and environmental policy through machine learning* (Research Project). researchgate.net

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rivas, A. (2019). The current research landscape of the application of artificial intelligence in healthcare. *Journal of Health and Technology*, *9*(4), 451-460.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education Limited.

Sabherwal, R., & Grover, V. (2024). The societal impacts of generative artificial intelligence: A balanced perspective. *Journal of the Association for Information Systems,* 25(1), 329–341.

Selbst, A. D., & Selbst, S. W. (2019). Invention of fairness. *University of Pennsylvania Law Review*, *167*(3), 697–767.

Sharkey, N. (2012). *Swarm robotics: A new perspective for intelligence in groups*. Springer Science & Business Media.

Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. *Proceedings of the 2015 ACM SIGSAC Conference on Computer and Communications Security*, 1310–1321.

Solove, D. J. (2008). The future of privacy. *Survival & Liberty*, *39*, 14–20.

Sparrow, R. (2016). *Understanding agency: For human flourishing*. Oxford University Press.

Sunstein, C. R. (2020). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30, 5998-6008.*

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from*

wrong. Tim Duggan Books.

Wikipedia contributors. (2023). Deepfake. In *Wikipedia, The Free Encyclopedia*.

ZOOZ Consulting. (2024). *The impact of generative AI on education and personalized learning experiences*. ZOOZ Strategic Insights. http://zooz-consulting.com

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

ChatGPT 時代下人工智慧與公共政策的倫理考量：一個人本主義的視角[2]

魯俊孟*

摘要

本研究旨在剖析生成式人工智慧（GAI）在公共政策領域的倫理挑戰，並從人本主義視角建構治理框架，以平衡技術創新與社會風險。透過文獻分析與跨領域案例研究，發現 GAI 存在偏見、資訊誠信、隱私、透明度及失業等系統性風險，並提出在核心領域實施「真人主導」（Human-in-the-Loop）與分級管理之必要。研究結論強調：應確立個人自主、公平、透明、隱私與問責五大支柱，採取多管齊下的政策策略，以確保技術進化。符合人類價值。

關鍵詞： 生成式人工智慧（GAI）、倫理意涵、公共政策、偏見與歧視、倫理指南

*東海大學行政管理暨政策學系專任副教授。E-mail: lucg@thu.edu.tw。

[2] 論文發表於《2025 美國公共行政年會》，美國華盛頓特區，20025 年 3 月 28 日至 4 月 1 日。收件：2026/03/15， 同意刊登：2026/05/18。