Framework for Generative Intelligence Deconstruction & Governance, perspective from Non-Neutrality Technological

Peng (Helium) Bob
School of Management and Economics, Chinese University of Hong Kong (Shenzhen), Shenzhen, Guangzhou 518100, China

Abstract:

The rapid development of large language models (LLMs) benefits society in many aspects, but also poses ethical issues and governance challenges. This paper thoroughly explores the essence of LLMs and the root causes of ethical issues. It also unearths the ethical problems of LLMs that differ from traditional productivity tools. Based on the shortcomings of existing ethical governance models, this paper attempts to propose a lifetime-based LLMs ethical governance framework, focusing on the training, fine-tuning, and use stages of LLMs. In addition, a framework is constructed that includes self-correction, multi-guardian, and collaborative governance to achieve more effective and comprehensive governance. This paper aims to provide important theoretical and practical guidance for the ethical governance of LLMs, promoting the healthy development of LLMs.

Keywords: Machine Ethic, Large Language Model, Technology Governance

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

1. Introduction:

Artificial intelligence technology is rapidly reshaping human society and production lifestyles at an unprecedented pace. Representative large language models (LLMs), such as the ChatGPT series by OpenAI, Sora and Claude series by Anthropic, Gemini by Google, and ERNIE series by Baidu, represent the latest development in the field of artificial intelligence. These models, with their triple drive of massive data, powerful computation capabilities, and advanced algorithms, demonstrate outstanding abilities in simulating human cognition and decision-making processes. Since the public release of ChatGPT in late 2022, the field of large language models has gradually evolved into a technology development and application deployment system centered on OpenAI. Particularly in February 2024, with the release of the Sora video generation model by OpenAI, the field of large model technology has made a significant shift from text-based intelligent generation to video-based intelligent generation, achieving a leap from static to dynamic display, making Sora a multimodal simulator that connects the physical and digital worlds.

Compared to other applications of artificial intelligence, large models have made significant breakthroughs and fundamental innovations in terms of technological advancements. In addition to their extensive applicability, multi-modal capabilities, and abilities to generate intelligence, large models deeply integrate with various industries. This has resulted in fundamental changes in production models, technological innovation paradigms, content creation methods, and human-machine interaction patterns (Hadi et al., 2023). For instance, in the fields of science research, biological medicine, and software development, production approaches have shifted from the traditional fragmented and detail-oriented manual operation model to an integrated, efficient, and intelligent generation model based on platforms(Hou et al., 2024; Thirunavukarasu et al., 2023; Xi et al., 2023).

However, the application of large-scale artificial intelligence models may generate a series of ethical issues. For instance, on March 31, 2023, ChatGPT was banned in Italy due to issues related to user privacy leaks, while it was being investigated for suspected violations of data collection rules (GPDP, 2023). On January 29, 2024, the Italian Data Protection Authority (Garante per la protezione dei dati personali, GPDP) announced that OpenAI had violated the EU General Data Protection Regulation, which could result in a fine of up to 4% of global turnover (GPDP, 2024). Additionally, the ethical issues associated with large-scale models include algorithmic opacity, digital divides, and privacy leaks (Weidinger et al., 2021). These also give rise to new ethical issues such as model black-box effects, data copyright infringement, the further development of deepfakes technologies, the impact on human subjecthood, and the exacerbation of social stratification (Yan et al., 2024).

The issue of how to effectively prevent and manage ethical lapses induced by artificial intelligence-generated content (AIGC) has become a globally significant topic of concern. Concepts such as "Technologies for Good", "Responsible Research and Innovation", and "Machine Ethics" are widely advocated and practiced. International organizations, countries, and regions are actively formulating related ethical regulatory policies to balance management and technological innovation. Representative regulatory policies from international organizations include UNESCO's (2021) "Recommendation on the Ethics of Artificial Intelligence," the G20's (2019) "G20 Artificial

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

Intelligence Principles," the IEEE's (2019) "Global Initiative on Ethics of Autonomous and Intelligent Systems," the G7's (2023) "International Code of Conduct for Organizations Developing Advanced AI Systems," the EU's (2023) "Artificial Intelligence Act" and the accompanying "AI Liability Directive," the "Bletchley Declaration" jointly signed by 28 countries and the EU (2023), and the "Guidelines for Secure AI System Development" jointly issued by 18 countries including the UK and the US (2023). These policies aim to strengthen the ethical governance of artificial intelligence at both formal and informal institutional levels.

At the level of individual countries or regions, the ethical regulatory model in the United States is gradually shifting from a relatively lax to a stricter one. This shift can be discerned from the "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" executive order signed by President Biden (2023). As the largest developing country, China places significant emphasis on the ethical governance of large model-related issues, issuing a series of policy documents such as the "Ethics Guidelines for the Development and Application of Artificial Intelligence in the New Era" and the "Technical Ethical Review Measures (Trial)" as well as the "Interim Measures for the Management of Generative Artificial Intelligence Services" (2023). These policies gradually strengthen the institutional construction of ethical governance and establish a National Science and Technology Ethical Committee to promote science and technology ethical governance. The interim management measures for generative artificial intelligence services proposed in the China Artificial Intelligence Development Plan specify a regulatory approach that combines inclusiveness, caution, and classification and grading. It also clearly defines the responsibilities of service providers and end users. To enhance the legalization level of ethical governance of large language models, the 2023 legislative agenda of the State Council includes related laws on artificial intelligence and network data security. Countries such as the UK and Singapore have also issued corresponding regulatory policies to prevent and manage ethical deviations in large language models. Additionally, to strengthen ethical governance of large language models, various countries or regions have established specialized regulatory agencies. For instance, the EU Artificial Intelligence Act proposes the establishment of the European AI Office to enhance the transparency of general artificial intelligence models and control systemic risks.

However, the current ethical governance modes based on single elements and single processes, top-down approaches, and a results-driven approach are insufficient in effectively addressing and managing the ethical deviations caused by generative AI applications such as large models. To strengthen the ethical governance of large models, it is crucial to investigate the origins of these ethical deviations, conduct an empirical study of the identified ethical deviations, and systematically analyze the existing ethical governance paradigms. By identifying areas where current theoretical research is weak, governance systems need improvement, and governance practices need innovation, we can search for more appropriate and effective ethical governance paradigms.

Given the aforementioned background, the aim of this study is to delve deeply into the attributes of large models and the root causes of their ethical misconduct. By analyzing the core factors of large model ethical misconduct, we aim to comprehensively summarize the main manifestations of large model ethical misconduct and the challenges it poses to existing governance systems. On this basis, we propose a novel framework for large model ethical governance and corresponding governance ecosystems. This study aims to conduct a theoretical analysis of the root causes of large model ethical misconduct from a governance framework perspective, propose innovative solutions, and construct a governance ecosystem from a governance ecosystem perspective. This is all aimed at achieving the effective reconstruction of large model ethical governance order.

This study presents the content as follows: Firstly, this paper unpacks the characteristics and key components of large-scale generative artificial intelligence models, thereby reinterpreting their essence from a profound perspective. Secondly, this paper conducts theoretical analysis on the root causes of ethical violations in large models, deeply examining the specific roles played by core elements such as algorithms, data, and humans in the ethical violations of large models. Thirdly, this paper comprehensively summarizes and summarizes the phenomenon of ethical violations in large models, and further expands on existing ethical violations in artificial intelligence or risks. Finally, this paper constructs a life cycle-based ethical governance framework for large models, with key components at its core, and establishes an ecosystem for large model ethical governance, effectively preventing and addressing ethical violations in large models.

2. Literature Review

The study of artificial intelligence ethics originated from the initial exploration of the relationship between ethics and digital technology in the 1940s and 1950s (Siau & Wang, 2020; Stahl, 2021). As technology advances, the field has gradually deepened and expanded. The scope of ethical governance has evolved from technology ethics to artificial intelligence ethics, eventually encompassing the ethical governance of large-scale generative AI models. Particularly after the emergence of ChatGPT in December 2022, a significant increase in research on the ethical governance of large models began. The literature related to this paper encompasses the following three areas: 1. Research on the societal and economic impacts of large models and their ethical imbalances. 2. Discussions on governance models, governance goals, and their implementation pathways, regarding the ethical governance of large models. 3. Research on ethical governance centered on key elements such as algorithms, data resources, and deep synthesis technologies, related to large models.

2.1. Research on the Social and Economic Impacts of LLMs and Their Ethical Misconduct

With the continuous advancement of artificial intelligence technologies, such as generative artificial intelligence (GAIs), these technologies have played a significant role in driving economic growth, enhancing production efficiency, and optimizing resource distribution. This has effectively facilitated the improvement of quality, increase in efficiency, and deepening of structural reforms in the supply side (Alto, 2023; Eloundou et al., 2023). However, at the same time, the application of these technologies also presents a series of ethical issues. These include privacy breaches, deepening digital divide, biased output, and fuzzy attribution of intellectual property rights, among others. Weidinger (2021) and others have categorized these issues into two main categories: technical-

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

inherent ethical risks and technical-application ethical risks. In different application fields, the application of LLMs may trigger various ethical imbalances.

The widespread application of large language models (LLMs) in research fields such as medicine, economics, and finance portends significant paradigm shifts in scientific research. These models can independently propose research hypotheses, design and implement scientific experiments, and validate the rationality of the hypothesis, marking a transition from traditional low-efficiency workshop-style research models to high-efficiency platform-style research models. This transformation has a profound impact on the breadth of scientific innovation (Liu et al., 2023; Wang et al., 2024). In addition, large models can provide auxiliary research services such as literature reviews and data acquisition for researchers, yet this could also lead to a series of ethical imbalances due to data copyright infringement, the difficulty in defining responsibility, intellectual property infringement, and scientific misconduct (Karamolegkou et al., 2023; Meeus et al., 2024). Especially, the blurring of responsibility definition will further exacerbate the chaos in the social trust system. Large model applications are also significant in industries such as industrial design, drug development, and materials science, and their replacement of simple repetitive labor may trigger issues such as social division or discrimination (Yilmaz et al., 2023).

2.2. Research on the Ethical Governance of LLMs

In response to the ethical misconduct issues and new challenges posed by large-scale models, the academic community has conducted in-depth research on the ethical governance of large-scale models. McLennan (2022) and others proposed that LLMs should undergo ethical design to reduce the probability of ethical misconduct. Based on the existing knowledge and input information, large models can generate content that is both valuable for thought and fully embodies moral connotations through the constructed mental framework. One effective governance method for the ethical behaviors of large models is to introduce ethical experts during the development stage. Furthermore, strategies such as constructing an ethical database or knowledge base, embedding values, designing for sensitive values, prompt engineering, and incorporating built-in correction mechanisms are also viable. These approaches enable large-scale generative AI to grasp the human moral ethical knowledge system, thereby reducing the likelihood of ethical misconduct.

Research on the ethical governance systems or frameworks in generative artificial intelligence, general artificial intelligence, and generalized models has emerged as a focal issue in another field of ethics (de Almeida et al., 2021). In the existing literature, models of ethics governance in artificial intelligence have gradually matured. These models mainly include regulatory governance, innovative governance, autonomous governance, market-oriented governance, centralized governance, agile governance, complex systemic governance, and resilient governance along the industrial chain, among others (Larsson, 2020; Xue & Pang, 2022). These governance models provide diverse governance paths and theoretical foundations for the ethical standards in the artificial intelligence field, providing important guidance for its sustainable development.

Jakob et al. (2023; 2024) propose that a hierarchical governance system should be established for GAI, from model training to implementation, to achieve comprehensive governance of complex systems. Related studies emphasize that in relation to the characteristics of large-scale models, governance strategies should shift from controlism to pedagogy. This governance paradigm should include the following three basic elements of law: the motivation for the development and training of large-scale models, the ecological construction of training data content, and the normative institutional system for specific scenarios and users.

The goals, pathways, and legal status of large model ethics governance vary depending on the specific application scenarios, showcasing a diversity of approaches. Hacker et al. (2023) conducted a deep analysis of the key dimensions of generation and removal, emphasizing the timeliness and adjustability of the liability determination standards. Meanwhile, Novelli et al. (2024) point out that the current ethics governance policies face numerous challenges, such as operational deficiencies, insufficient constraints, and limitations in terms of effectiveness. They advocate for prioritizing legal principles and accelerating the construction of a standardized system for ethical governance of technology. However, some scholars hold differing views, stating that the large model technology is still in the early development stage and has not fully realized its practical applications. Initiating targeted ethics governance and legislation at this stage may not be advisable, as premature intervention or regulation could hinder the innovation and development of technology, leading to deviations from the original functionality of legal norms. Therefore, King et al. (2024) suggest that governance should be achieved through the establishment of technical standards, risk assessment of security, and improvement of information review technologies.

2.3. Research on the Ethical Governance of Key Elements in LLMs

The relevant research on the ethical governance of LLMs primarily focuses on two key aspects: data and algorithm. LLMs require a massive amount of data for training, and the size of the data has a highly positive correlation with the ethical knowledge that these large models can learn and extract from humans (Krijger et al., 2023). However, the ethical risks associated with misleading, biased, and discriminatory training data cannot be overlooked (Boyd, 2021). To strengthen the ethical governance of data, Hosseini et al. (2022) argue that the ethical issues of data in the information age do not only require the audit work of data providers, but that the non-formal governance mechanisms formed by individual, organization, and societal interactions should also be a part of the governance framework. In addressing issues such as data quality, security, and timeliness, Abraham (2019) et al. propose constructing a precise and multi-dimensional data accountability matrix, to build a flexible and efficient data governance regulatory system.

Addressing ethical issues such as algorithmic black boxes, algorithmic biases, and algorithmic discrimination, Zorrilla et al. (2022) propose a scenario-specific algorithm regulatory approach, constructing specific institutional mechanisms for algorithm openness, data anonymization, and anti-algorithmic discrimination to achieve algorithmic accountability. In response to algorithmic ethics risk issues, Martin (2019) proposes a framework for dynamic regulation throughout the entire process to achieve agile governance. However, the principles for allocating responsibility among different stakeholders and for users' algorithmic responsibility are still unclear (2019). Furthermore, Li et al. (2024) propose a comprehensive governance of large models from the perspective of deep synthesis technology ethics, aiming to establish a more influential governance legal system worldwide.

Through a deep analysis of existing literature, the primary deficiencies in current research are as follows: Firstly, although there has been independent research on the ethical governance of key elements such as data and algorithms, there is a lack of specificity and adaptability in terms of the ethical governance of key elements in large-scale models. Secondly, existing research fails to detail the ethical governance of key elements at different stages of the life cycle of large-scale models, and lacks a systematic analysis of the causes of ethical misconduct in these key elements. Additionally, there is a lack of corresponding ethical governance frameworks. Given the complexity of large model technology, its uncertain outcomes, and its broad applicability, these risks manifest in a wide impact range, rapid propagation speed, and strong multiplier effects. The existing ethical governance models are clearly insufficient to effectively address these challenges, necessitating the development of new ethical governance frameworks to address these problems.

3. Deconstruction of Ethical Misconduct in LLMs

Compared to other artificial intelligence technologies and applications, the exploration of the root causes of ethical violations in LLMs necessitates a focus on their essential properties. This study draws on Marx's perspective on the relationship between labor productivity and production tools, dissecting the essential properties of large models. By analyzing the characteristics of large models at various key stages of their life cycle, this paper aims to explore the origins of the attributes of large models and the ethical violations they exhibit, providing theoretical grounds and path choices for effectively preventing and managing the ethical violations of large models.

3.1. Origins of Ethical Misconduct

3.1.1. Productivity Attributes: Technological Non-Neutrality

LLMs represent advanced progress in the field of natural language processing (NLP), which have been developed through the integration of algorithms and frameworks on the basis of powerful computational infrastructure and large-scale datasets. Transformer-based neural network models have evolved into three main architectural patterns: masked language models, autoregressive language models, and sequence-to-sequence models. LLMs as a sub-category of artificial intelligence technology fall under the productivity category, and as a new type of productivity centered on technological innovation, they represent the most active driving force for social progress, signifying a significant improvement in societal productivity. (Nakavachara et al., 2024). Marx views productivity from a dual perspective: physical productivity is the objective strength of human beings in material production, while mental productivity is the ability of human beings to produce spiritual wealth. The unity of the two lies in the fact that both physical and mental productivity have generated use value. Hence, the generative content generated by LLMs primarily satisfies human mental needs, and they fall under the category of mental productivity.

The core technical elements of large models are algorithms, which also fall within the scope of productivity. They possess black-box characteristics. Kowalski(1979) defined algorithms as a series of computational rules or steps. Algorithm black boxes can be divided into subjective black boxes and objective black boxes, which further lead to phenomena of ethical deviation such as algorithmic bias and algorithmic discrimination (Mühlbacher et al., 2014).

The generation of the "algorithm black box" phenomenon is primarily attributable to three aspects: Firstly, from the perspective of intellectual property, algorithms themselves possess a distinct commercial secrecy and economic value, which makes it impossible to disclose them to the public. For instance, despite the replicability of the GPT-4 model by OpenAI, the company is unwilling to disclose it in the short term. Secondly, as a technology, the inherent opacity and inexplicability of algorithms prevent human beings from predicting the results they generate. Finally, the "algorithm black box" phenomenon is also reflected in the cognitive biases of different groups towards the operational mechanisms and decision-making processes of algorithms. This bias is particularly evident between professionals and non-professionals. In the ethical risks associated with large models, the "algorithm black box" manifests itself through users' inability to observe or verify the data processing process of the algorithms, leading to uncertainty and inexplicability in the results. Therefore, from the perspective of productivity theory, large models no longer possess the "value neutrality" characteristic and instead are accompanied by the "algorithm black box" phenomenon and related ethical violations.

From the perspective of the labor object, large models that embed human ethical tendencies or preferences cannot maintain neutrality. Compared to other forms of productivity, the labor objects that large models operate on are data. With characteristics such as virtuality, repeatability, and the inclusion of human ethical and moral content, data represents a new form of productive factor. Its storage or management requires relying on networks and other carriers, which adds a virtual quality to it. Data's characteristics that allow for its repeatable use and processing, as well as the difficulty in controlling its future flow and usage path, expose them to specific data ethical risks. Additionally, data elements inherently embody human values, ethical tendencies, or preferences, especially personal data, further endowing them with a "personified" quality. Since the data elements that large models operate on not only include human ethical tendencies or preferences but also are themselves potentially susceptible to ethical risks, these data elements, when incorporated into the operation of large models and embedded in the generated content, may further propagate and amplify ethical preferences, tendencies, and risks.

3.1.2. Embedding Human Ethics

LLMs, as applied systems, serve as intermediaries between humans and the processing and transformation of input content. They can be considered an extension of human brain functions. Therefore, large models are integrated into the category of production tools. During the development phase of large models, developers incorporate their own ethical beliefs and preferences into the system. In the application process, the ethical content generated by the large model may potentially have profound ethical implications on human society.

During the development stage of large-scale models, the ethical preferences of the developers will inevitably be embedded in the models, leading to the "black box" phenomenon of algorithms. With the continuous advancement of AI technology, these LLMs are attempting to emulate the operation mechanisms of human brains, aiming to achieve parameter numbers comparable to those of human brain synapses. For example, GPT-4, a large model, is composed of eight different mixed models, each containing approximately 220 billion parameters, with a total of over 1.8 trillion parameters, which is ten times more than the parameters of its predecessor, GPT-3. In this process, the selection and setting of algorithm models primarily rely on the subjective preferences of the developers. Therefore, as LLMs development technology advances, the operation mechanisms of these models will increasingly resemble those of human brains, with the ethical biases and preferences of the developers embedded in the model. This will make the transparency of the algorithm an invisible companion throughout the entire life cycle of the model.

The ethical biases or inclinations of the training data, trainers, and data processors may be incorporated into the process. The origins and quality of the training data themselves may already contain ethical biases or inclinations. To optimize the quality of LLMs' training data and reduce their dependence on data size, quality screening, removal of redundant information, and elimination of privacy content are necessary. However, in this process, the ethical preferences or biases of human trainers or data processors may be embedded. Due to the vastness and complexity of the training data and its content, data processors may not be able to handle them accurately, thus exposing ethical risks. Compared to original data, the fundamental difference of synthetic data lies in the fact that besides the ethical inclinations or preferences inherent in the data itself, the data synthesis process may also implant the ethical inclinations or preferences of data processors, thereby influencing the fairness of its content. The processing and annotation methods of training data also encompass ethical biases or preferences of human trainers or data processors. The primary source of large model pre-training data is original data, which is supplemented by synthetic data, while the model tuning data is primarily based on annotation data. In this process, code-type data mainly comes from public code repositories on GitHub or Huggingface, with Python files predominating; for dialogue-type data, it primarily relies on annotated data, with professional annotators and users of large models constituting the annotators. Hence, the large models inevitably encounter ethical risks and limitations in their training data processing and annotation due to the ethical inclinations or preferences of data handlers, data annotators, and users. This can lead to the exposure of ethical risks and the limitations in making inferences based on biases.

As generative models, LLMs not only extend and replace human physical labor, but also represent the extension and replacement of human brains and nervous systems. The emergence of large-scale models represents a revolutionary shift in the paradigm of production tool supply. They extend and replace human physical labor. Human beings are naturally limited by their physiological conditions in terms of their ability to labor under certain environmental conditions. For instance, they cannot sustain productive activities in extreme temperatures or low temperatures. In order to overcome these physiological limitations, production tools have emerged as bridges connecting workers and production objects. These production tools are not only tools utilized by humans in labor, but they also serve as key markers for delineating different historical periods in human society.

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

As key components of the industrial society, machines play a pivotal role. In the digital economy era, intelligent machines empowered by artificial intelligence technology, such as intelligent assistants or robots, represent a significant form of their manifestation. Unlike traditional production tools, LLMs primarily operate on data elements during their application process, producing content with prominent virtual and digital characteristics. The emergence phenomenon of intelligence in LLMs, and their similarity to human thinking, have led to a paradigm shift in the supply of production tools. LLMs are viewed as the extension and replacement of human brains and nervous systems, and their application procedures actively incorporate human ethical inclinations or preferences. However, the development of LLMs poses unprecedented challenges to human subjectivity. Specifically, the mass production of machines is replacing traditional handcrafting, leading to an increasing dependence on repetitive and mechanical work, further exacerbating the phenomenon of human objectification and alienation. In this machine-centered production system, humans are placed in mechanized production lines, their labor intensity increases, and they are subject to both mental and physical stress, limiting the possibility of their full development.

In the era of digital economy, the application of GAI is ubiquitous, permeating deeply into people's production and daily lives. The relationship between humans and machines is increasingly close, moving from machines serving humans, to human-machine collaboration, to symbiosis, and ultimately to integration. This leads to a gradual blurring of the boundary between humans and machines, potentially leading to an over-reliance on GAI and the formation of 'proxies', even turning humans into 'technological dependents' (Pani et al., 2024). This trend poses a serious threat to human agency and the loss of decision-making autonomy.

3.1.3. Dual Attributes: Fusion of Human Ethics and Machine

LLMs (large language models) are comprehensive systems that integrate algorithms and models in artificial intelligence technology, closely connected with applications, forming innovative digital productivity tools. OpenAI's ChatGPT series and Sora are examples of this, where their foundational technologies respectively adopt the Transformer architecture and Diffusion Transformer architecture. Their application layers are realized through GPT-4 and Sora large models. Large models are not only a synthesis of productivity and production tools, but also possess these features. Throughout their development process from research to application, they incorporate human ethical preferences. The generated content naturally carries strong ethical bias. Karl Popper believed that the method of science is the method of bold conjectures and ingenious and severe attempts to refute them. As per Popper's view, large models can be considered a symbiotic combination of science and practice. They extend the human mind or consciousness into machines. Unlike brain-computer interfaces that seek to directly acquire and decode human thoughts, large models learn human thoughts or consciousness, demonstrating emergent capabilities. This is a characteristic of continuous human-machine integration evolution, and an external manifestation of machines constantly mixing with human ethical and moral systems. Hence, as a digital productivity tool, large models possess both the dual attributes of productivity and production tools. The ethical violations at the source or basic logic of these models are the result of the comprehensive fusion of these two aspects. These ethical violations are pervasive throughout the entire lifecycle of large models (Table 1), which makes them more complex and challenging to manage.

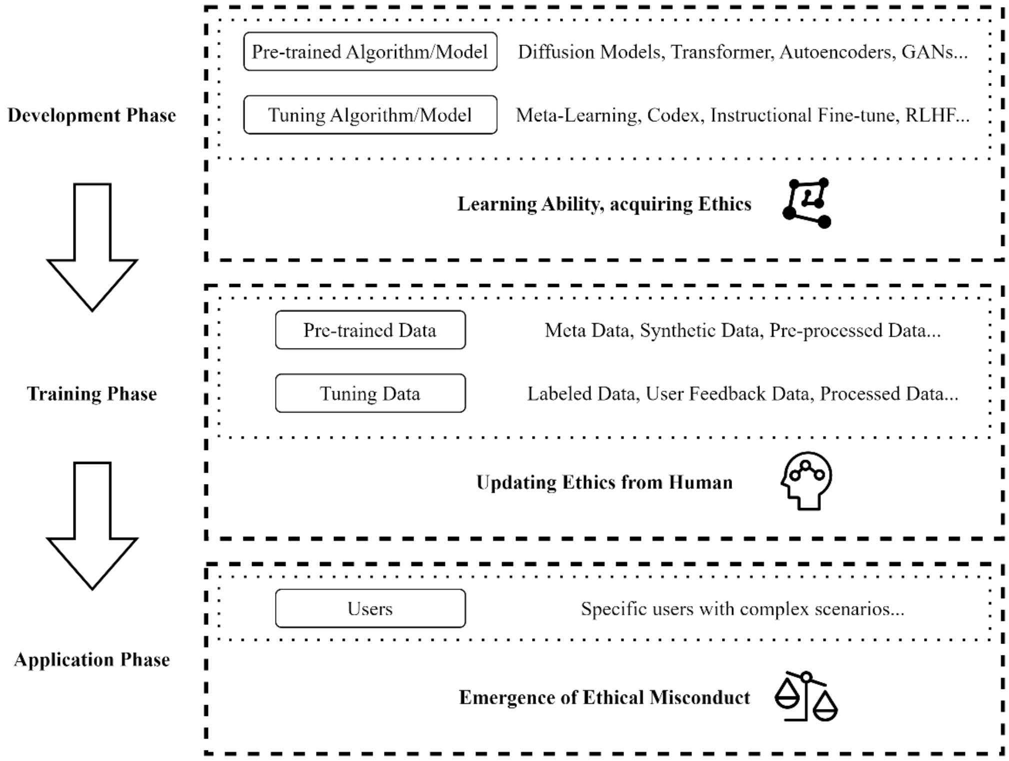
Table 1 The origin of ethical misconduct throughout the entire lifecycle of models

	Factors	Actors	Phenomena
Development Phase	Algorithms Models	Related Developers	Algorithm Black Box
Training Phase	Training Data Fine-Tuning Data	Data Processors Data Annotators Model Testers	Data Security Privacy Invasion Copyright Infringement
Application Phase	User Data Generated Content	Service Providers Application Users	Algorithmic Bias Deep Fakes Digital Divide Social Stratification

3.2. Deconstruction of Ethical Misconduct in LLMs

By conducting a three-dimensional analysis on productivity, production tools, and their composite, we can reveal that the three key elements of algorithms, data, and human users are the primary factors leading to the ethical deviation of large models. Among them, algorithms constitute the technical basis for ethical deviation; data plays a mediating role; and human users directly drive the phenomenon of ethical deviation in large models. The relationship between the life cycle of large models, their key elements, and ethical deviation is illustrated in Figure 1.

Figure 1 Relationship between the lifecycle of large models and ethical misconduct


Development Phase
Pre-trained Algorithm/Model
Diffusion Models, Transformer, Autoencoders, GANs...
Tuning Algorithm/Model
Meta-Learning, Codex, Instructional Fine-tune, RLHF...
Learning Ability, acquiring Ethics
Training Phase
Pre-trained Data
Meta Data, Synthetic Data, Pre-processed Data...
Tuning Data
Labeled Data, User Feedback Data, Processed Data...
Updating Ethics from Human
Application Phase
Users
Specific users with complex scenarios...
Emergence of Ethical Misconduct

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

3.2.1. Technological Origins: Algorithms

Algorithms form the core logic of large-scale model operations, producing content that aligns with human expectations through processing and learning from input training data. However, the "black box" problem extends further to the output content, manifested in its untraceable origin and inherent uncertainty. Moreover, as algorithms learn from data and are widely deployed across various application scenarios, they gradually integrate into various aspects of human life. In this process, algorithms that were initially deemed "value neutral" have become learners, reflectors, and influencers of human local moral and ethical systems. If the "black box" of algorithms is transformed into a "transparent box", making the internal logic of its operation clear, then the content generated by large models can be predicted, allowing the root causes of ethical violations to be traced and corrected. Thus, algorithms themselves serve as the technological root of model ethical violations, while simultaneously being the source of a series of ethical violations such as algorithm "black box" phenomenon, algorithmic discrimination, algorithmic bias, and the model "black box" phenomenon, model discrimination, and model bias.

3.2.2. Key Intermediary: Data

The rise of large models signifies a fundamental shift from traditional knowledge generation models to machine-driven generation models. These models leverage similar mechanisms to the human brain, achieving intelligent breakthroughs in content generation in a short period. The training and operation processes of large models deeply depend on massive amounts of data. Simultaneously, they absorb and inherit the ethical moral systems inherent within these data during the learning process. Through a "mirror effect," the content generated by these models reflects local human ethical moral systems, thereby exerting influence on humans' ethical moral systems (Hagendorff & Danks, 2023; Zhao et al., 2021). Based on the differences in the goals of the training data and the development sequence of the ethical moral systems formed by LLMs, the ethical moral systems can be divided into primary ethical moral systems and derivative ethical moral systems. The primary ethical moral system refers to the initial ethical moral framework formed by the researchers, trainers, data annotators, and other participants in the model's development, training, and optimization. The derivative ethical moral system is formed when large models enter the practical application stage. The application users incorporate their own ethical moral views or the ethical moral views embedded in the input content into the model, which, in the process of intelligent interaction, influence the primary ethical moral system, thereby establishing a new ethical moral framework or causing the evolution of the existing ethical moral system. The ethical and moral framework of LLMs encompasses the ethical and moral systems of all stages of participants and the input data. Within this framework, input data is treated as an objective entity, reflecting the ethical and moral perspectives of the aforementioned related entities. Therefore, the ethical and moral system constructed by large models is actually a complex system formed by integrating the ethical and moral perspectives of multiple human entities through artificial intelligence techniques, such as algorithmic processing.

3.2.3. Behavior Subject: Humans: Human

The actors involved in unethical behavior of large models, the affected entities, and the responsible entities are all human beings. Under this framework, algorithms act as proxies on behalf of human beings, and large models also play the role of human agents. The distinction lies in the scope and intensity of the agent's content. Data, as the product and resource of human activities, reflects human ethical and moral systems. From this perspective, data itself possesses a "value neutrality," and whether it triggers ethical misconduct depends on the production mode, application methods, and specific application scenarios of humans. In summary, although large models pose challenges to human subjectivity, humans still control large models and determine their content. Large models reflect human will, regardless of the ethical responsibilities they bear. Ultimately, the related ethical responsibilities fall to humans.

3.3. Manifestations of Ethical Misconduct in LLMs

The ethical transgression of large models primarily stems from their characteristics such as algorithm stacking, parameter explosion, diversified training data, and scale growth, and the strengthening of intelligent emergence capabilities. This leads to the emergence of new ethical transgressions and the deepening or expansion of existing issues. For instance, model black-boxization, data copyright infringement, difficulty in determining responsibility, challenges to the human subject's status, and disregard and destruction of social moral and ethical order.

In the field of artificial intelligence and algorithmic ethics research, the algorithmic black-box problem is a major challenge to ethical governance, involving topics such as algorithmic discrimination, bias, and uncertainty in results (Rai, 2020). With the application of deep neural networks and integrated technologies, the algorithmic black-box gradually evolved into the model black box. Algorithms, as processing rules for data, have deeply permeated social life and influence value judgments, losing their "value neutrality". Models are output files derived from data training by algorithms, containing specific processes or structures, algorithms, and parameters. The parameters in the model influence the output through training and optimization, with the number of parameters affecting the accuracy of the output. The model black box is more complex than the algorithmic black box, which is the result of the entanglement of algorithms, models, and parameters. It exacerbates the interpretability and uncertainty of large-scale model content, and its generated content may be incorrect, false, harmful, or misleading, lacking causal inference. The phenomenon of machine illusion can produce negative impacts (Ji et al., 2023). Compared to specialized artificial intelligence models, large models exhibit a stronger emergence of intelligence and faster generation speed, making it easier to generate and rapidly disseminate low-quality content, potentially posing systemic challenges to social moral and ethical order.

With the advancement of information technology, the problem of data copyright infringement has become increasingly severe, and the identification of the responsible entity has become increasingly difficult. Data and text mining technologies have made it easier to access and use public data. Large language models rely on massive amounts of data for their intelligence, but their legitimate use of data and text needs to be strengthened by regulations. The Directive on Copyright of the Digital Single Market of the European Council(2019) regulates the acquisition and use of data and text, allowing for the unlicensed use of data without copyright protection, but also requires the prevention of copyright infringement. Large models are partially trained using data from network crawling, but the legal and security aspects of data use have not been assessed, potentially leading to the violation of data subjects' rights and the leakage of sensitive information. The ambiguity of the human-machine boundary, data source, and data subject rights makes it challenging to identify the responsible entity. The same issues arise when using synthesized data to train large language models. The emergence of Sora video models extends the problem of data copyright to the field of video generation.

The blurring of the human-machine boundary, especially between the physical and virtual worlds, challenges the status of humans and the social trust structure. In the digital economy era, the development of large model technologies such as ChatGPT and Sora, which has disrupted the binary structure of production relations proposed in political economy, has led to a significant escalation in human alienation. This technological progress not only blurs the human-machine boundary but also weakens the agency and decision-making autonomy of humans, affecting significantly the human subjectivity. The introduction of Sora, a video large model, has exacerbated the problem of deepfake. Deepfake, an AI technology widely used in video face replacement, is based on generative adversarial network models and has high realism, broad applicability, and rapid development. This may lead to ethical issues such as infringement of individual rights and disruption of social trust, among others. Compared to deepfake, the substantial breakthrough of Sora's video large model exposes the depth of forgery to a new level, moving from the forgery of faces to the forgery of the physical world, further threatening the social trust system.

Finally, the outputs of large-scale models pose a disregard and disruption to the moral and ethical order of human society. In particular, when applied to specific use cases, the large-scale models carry along the local moral and ethical systems that they have learned, generating content for the application users and blending in the original moral and ethical systems with the derivative ones. The developers, service providers, and users of large-scale models disregard whether the generated content disrupts the moral and ethical order of human society. The generated content of the large-scale models exists in the moral and ethical vacuum zone of human society, untainted or unregulated. For instance, the use of artificial intelligence to generate deceptive content to interfere with elections is considered a major global challenge (Ferrara, 2024; Schmitt et al., 2024). The misuse of large-scale models disrupts the moral and ethical order of society. They serve as "human brain" or "world simulators" and shift the process of knowledge generation, reduce the necessary labor time, and reduce the cost of obtaining generated content. However, there are no effective governance mechanisms in place for the entire process from input needs to output results and their future use, which naturally raises questions about the moral and ethical order of society.

3.4. Governance Challenges by Ethical Misconduct

The technological revolution and the re-structuring of the current order brought about by large-scale models pose profound challenges to the current ethical governance system. The existing governance models, including those based on single elements and process steps, top-down ethical governance models, and models centered on outcomes, are severely inadequate in effectively preventing and responding to the deviations and violations in ethics manifested by large-scale models. This is particularly prominent in the absence or failure of ethical governance.

Ethical governance of a single element or phase is insufficient to effectively prevent ethical issues in large models. The blurring of the human-machine boundary and the reconfiguration of the human-machine relationship lead to an expansion of the ethical governance subject. With the evolution of human-machine intelligence interaction and feedback learning mechanisms, the reconfiguration of the human-machine relationship generates a "flywheel effect," resulting in a change in the scope of ethical governance (Gurkan & de Véricourt, 2022). This is reflected in the change of the ethical governance objects in large models. Traditional ethical governance primarily targets content and service providers, while the ethical risks brought about by large models encompass a wider range, including developers, trainers, users, as well as providers of training data, processors, and annotators.

There is a lack of clarity in the identification of responsible parties and responsibilities in large model ethical governance. For instance, the ethical risks associated with the generation of content are closely related to the ethical considerations during the development process. These responsibilities are shared by the developers, service providers, and users of the output. When ethical violations occur, it is challenging to delineate the responsibilities due to the opacity of the algorithms and the inherent uncertainty of the results. Therefore, traditional single-level governance models are ineffective in effectively preventing and handling large model ethical issues.

Top-down ethical governance primarily relies on regulations, but their delay and inflexibility hinder timely prevention and anticipation of potential ethical risks arising from the continuous technological iteration and updates of large model technologies. The uncertainty brought about by the generation of content and social implications based on the technology unpredictability of large models necessitates a non-linear, rigid governance model to effectively prevent their impact on the human society's moral and ethical order. A top-down ethical governance approach primarily driven by government regulations can lead to a disconnect between ethical standards and practices. Therefore, we need to shift from ethical descriptions to applications, explore novel governance models, and prevent the unchecked growth of large models.

The result-oriented ethical governance model does not perform well in preventing and addressing ethical issues in the life cycle of large models. This model primarily monitors the content or information generated by large models and their applications, and primarily assigns responsibility to generative AI service providers. However, the ethical deviations caused by the development and application patterns of large model technology are not limited to the generation of content. It is difficult for the result-oriented ethical governance model to effectively prevent ethical deviations merely by managing its ethical risks. The innovative paradigms of large models, such as open-source innovation, process innovation, and uncertainty innovation, make it impossible for the result-oriented ethical governance model to effectively manage its ethical risks. Simultaneously, as the application scenarios of large models become diversified and open, the potential boundaries of ethical deviations expand, exhibiting a multi-point, multiple-occurrence character. This makes it impossible for the result-oriented ethical governance model to effectively prevent and address the ethical issues of large models.

4. Reconstructing the Ethical Governance Framework

On the basis of ethical issues such as algorithmic black box and digital divide, large models have further generated phenomena of model black box, data copyright infringement, and unclear responsibility subjects, thereby threatening the human subjectivity. These ethical issues pose challenges to the existing ethical governance system. Currently, single-element, single-step, top-down, and outcome-oriented ethical governance models have exhibited issues of empty governance, delay, solidification, and policy and practice decoupling. Therefore, it is urgent to propose a new ethical governance framework to effectively prevent and respond to the ethical deviance caused by large models. This study examines the ethics governance challenges posed by the large models, based on the analysis of its properties, the root cause of ethics deviance, the dissection of key elements, and its main manifestations. Furthermore, it discusses in detail the governance challenges presented by the ethics deviance of large models. This study constructs a comprehensive life cycle ethics governance framework for large model key elements, anchored on the life cycle of large models and its core elements.

4.1. Fundamental Approaches

In the construction of the ethical governance framework for large-scale models across their entire lifecycle, this paper follows the following fundamental approach. Firstly, the governance framework aims to cover the entire lifecycle of large-scale models, from research and development, training, to practical applications, even reaching its withdrawal stage, to ensure the continuity and completeness of ethical governance. Based on this, the ethical governance of large-scale models should not only encompass all stages of their lifecycle, but also extend to the withdrawal stage. Secondly, the framework emphasizes the ethical governance of key elements of large-scale models, namely algorithms, data, and human beings as agents. By placing these key elements in specific application scenarios and emphasizing the cultivation of internal ethical literacy and self-regulation, to enhance the global applicability, adaptability, and self-implementation effects of ethical governance. Lastly, the framework aims to achieve alignment or ethical consistency between large-scale models and human values by establishing artificial moral agents.

The compatibility with human values refers to the fact that large models should align with human values and maintain consistency with human value systems and social and ethical systems. Human values encompass various dimensions such as survival, social, and political aspects. The social ethics system mainly involves formal and informal institutional structures within the scope of social ethics. The compatibility of large models with human values can be demonstrated from three perspectives: cultural adaptability, social adaptability, and legal compliance. Specifically, cultural adaptability requires the large model to conform to the cultural environment of the country or region where it is used. Social adaptability requires the large model to align with the social environment of the country or region where it is used. Compliance with laws entails compliance with national or regional legal and institutional systems. The first two emphasize compliance with informal institutions, while the latter emphasizes compliance with formal institutions.

It is crucial to emphasize the systematization and integrity of ethical governance in large-scale computational models. To ensure the effective implementation of the governance framework based on key elements of large-scale computational models throughout their entire lifecycle, establishing an ethical governance ecosystem is not only a refinement of this governance framework but also a crucial supplement. In this context, the dual-level actor and dual-level gatekeeper mechanism within the ethical governance ecosystem of large-scale computational models plays a significant role in analyzing the key elements of these models. Additionally, a global cooperation governance network extends the governance of key elements, providing a more effective approach. Specifically, through promoting the innovation of technologies related to large-scale computational models, we can attain precise governance of these models' algorithms and other technical dimensions. Meanwhile, international organizations and national governments play a pivotal role in global and cross-regional ethical governance. The social public opinion system also demonstrates its indispensable role in supplementing and supervising formal ethical governance mechanisms.

4.2. Ethical Governance

4.2.1. Ethical Governance towards Technology

Large models essentially represent the comprehensive embodiment of algorithms, models, data processing and annotation, and model training technologies. In large models, we need to conduct a comprehensive evaluation of the correlation and divisibility of the algorithm, model, and parameters. We then regard these three elements as important components of large model technology. One cannot overlook the importance of algorithm ethics governance in this context. With respect to the ethics governance of large model technology, a thorough exploration can be conducted from three perspectives: Algorithm Ethics, Model Ethics, and Data Ethics.

- The evolution of ethics in large model technology. Strengthening technological innovation in areas such as algorithms, model itself, model training and optimization, data processing and annotation. For example, enhancing transparency, interpretability, and ensuring ethical correctness in generated content by introducing RLHF (Human Feedback Reinforcement Learning), improving model interpretability, and operating model audits initiated by OpenAI (2022; 2023). Additionally, integrating ethical data training modules or built-in ethical self-check or audit mechanisms in large model frameworks can effectively reduce the occurrence and potential spread of ethical deviance.

- Ethical self-review, testing, and evaluation of LLMs technology. Considering the commercial secrecy and intellectual property value of LLMs technology, and in order to cover the key ethical risks that may exist in this technology as comprehensively as possible, we can draw on the model card system proposed by Mitchell et al. (2019). In this system, developers of large models are required to create model cards, which provide detailed records and documentation of the key information of the LLMs, such as parameters, algorithms, training datasets, expected application domains, and target user groups. This approach enhances the transparency of large models and strengthens their compliance with ethical norms.

- The technical ethical review and supervisory mechanisms of LLMs. Various countries and regions have implemented ethical governance measures such as registration, review, and safety assessment for artificial intelligence technology. To strengthen the rigor of external review on technical ethics for LLMs, a full-process ethical audit system can be considered. Appropriate audit methods can be selected according to different technical attributes, such as code auditing, non-invasive auditing, crowd-sourced auditing, agent auditing, and scraping auditing. This ensures the consistency of responsibilities between auditors and the audited entities, the independence of auditing institutions, and the full protection of intellectual property rights.

4.2.2. Ethical Governance towards Data

Data play a critical role in driving the operation of large models, which can be divided into two main categories: training data and user data. Among them, the training data of large models mainly encompasses textual, image, video, and data types corresponding to various formats such as text and video. These data types play a crucial role in the pre-training and fine-tuning stages of the large model, thereby affecting the ethical rules learned by the large model. As the applications of large models expand, the impact of these factors will reflect its extensiveness and expansiveness. Ethical management of large model data is a crucial constituent of achieving ethical governance of large models. Taking into account the ethical risks of large model data, this paper proposes to strengthen the ethical management of large model data from three aspects: data source, quality, and security.

For data originating from publicly available sources or professional databases that have not been processed, it is necessary to enhance the traceability, legitimacy, and compliance of their sources to prevent data copyright infringement and other ethical unethical behavior. Furthermore, the source of the data plays a significant, but often implicit, guarantee for its quality and the ethical values it carries. Data originating from legitimate sources tend to have higher quality and healthier ethical health.

The quality of data used by large models is reflected in whether it contains misleading, discriminatory, etc. These can impact the moral and ethical systems of large models, generated content, and end-users. It is crucial to enhance the detection and evaluation of training data quality and establish a corresponding process and methods for dealing with abnormal data. This ensures that the training data complies with ethical order and moral standards, preventing low-quality data from negatively impacting the moral and ethical systems of generated content and end-users.

Data security and privacy protection constitute the core underpinning of large model operations. Throughout the process of data storage and application, security requirements are ubiquitous, and it is imperative to accelerate the breakthroughs in key privacy computing technologies such as federated learning, multi-party secure computation, differential privacy, and homomorphic encryption. Simultaneously, it is also crucial to strengthen the interconnectivity between related fields, providing an innovative technology option for data security and privacy protection and reinforcing its security line. In terms of the management approach of data security, introducing specialized data hosting institutions can be a feasible solution by separating the responsibilities of data storage, management, and application, enabling effective monitoring of data processing and application, enhancing data security, and assisting government data management for large model training.

4.2.3. Ethical Governance towards Acting Subjects

Humans are the primary agents influencing the ethical and moral systems of large models. They are the core elements of their ethical management. At every stage of development, training, deployment, and application of large models, humans play a dominant role and are consistently involved. The foundation of ethical management in large models lies in the management of humans. Based on the different stages of the life cycle of large models, we can divide humans into the developers, trainers, data suppliers, data processors (mainly including data pre-processors, synthesizers, and labelers), large model service providers, and large model users. Ethical management of these behavior subjects requires a comprehensive consideration of their commonalities and individualities, i.e., the differences in the ethical risks exposed by different behavior subjects.

The ethical governance of large model behavior entities involves enhancing their ethical cognition level, ethical literacy, and professional skills. To fundamentally achieve ethical governance goals, it is essential to categorize and govern these entities in a stratified manner, elevating their ethical cognition level, and enhancing their capacity to mitigate ethical risks. In accordance with Ford & Richardson's (1994) four-stage theory of ethical decision-making - ethical cognition, ethical judgment, ethical intent, and ethical behavior - educational institutions can systematically educate these entities, public media can widely publicize ethics, and the workplace can continuously train them to elevate their ethical cognition and value judgment abilities, guide ethical judgment and intent, reduce the likelihood of ethical deviations, and diminish the negative impact of ethical deviance on human society's moral and ethical systems.

By enhancing the ethical literacy and professional technical skills of various behavior entities involved in large-model development, training, data provision, and management, as well as service provision, the likelihood of ethical misconduct can be effectively reduced. These behavior entities directly interact with and operate the algorithms, models, parameters, and training data for large models, gaining relative advantages in information acquisition and spatial advantage. This gives them a deeper influence on the "ethical and moral rules" that large models learn. Therefore, enhancing the ethical literacy and professional technical skills of these behavior entities is not only urgent, but also more significant for the effects of ethical governance. Professional ethics enhancement primarily focuses on forms such as on-the-job training and learning of professional ethics codes; professional skill enhancement mainly manifests in areas such as overcoming black box models, embedding ethics, and designing ethics, and integrating human society's moral and ethical rules into the entire process of large model development, training, and service provision, reducing the impact of large models on the human society moral and ethical order system.

External ethical governance of large model behavior. The implementation of licensure, market blacklist, and ethical misconduct accountability systems to strengthen the external ethical governance of large model behavior. Licensure is implemented before large models are introduced into the market, after ethical evaluations or performance tests performed by regulatory agencies. It allows large models that meet the standards to enter the application market. Market blacklist system determines whether to blacklist large model behavior entities based on factors such as the frequency of ethical misconduct, its scale, and the extent of damage or harm caused. Ethical misconduct accountability mainly involves punishing large model-related entities, compensating for the losses suffered by affected entities, and preventing potential ethical risks.

5. Ethical Governance Ecosystem Construction

To strengthen the ethical governance of LLMs and ensure their consistency with human values or ethical standards, it is essential to promote the participation of diverse stakeholders within the ethical governance framework for LLMs' key elements throughout their entire lifecycle. This system primarily involves promoting the self-governance of key groups and enterprises at both levels, while leveraging the gatekeeper roles of large platform companies and governments at the two levels. Sub-systems such as a global collaborative governance network centered around international organizations, guided by technical governance, and monitored by societal discourse, are established. This accelerates the "greening" process of the "deficit" in the ethical governance field of large models. The ethical governance ecosystem encompasses a diversity of stakeholders such as key groups, enterprises, digital gatekeepers, international organizations, and governments. These stakeholders participate in governance through mechanisms such as self-governance, mutual governance, and external regulation, achieving comprehensive coverage of factors influencing moral ethics. This system does not limit itself to considerations of a single factor or phase. It ensures comprehensive evaluation of ethical issues from multiple perspectives. The self-governance of key groups and the regulatory gatekeeper role of digital gatekeepers effectively supplement traditional top-down ethical governance models led by governments. They provide new perspectives and practical paths for constructing more comprehensive and diverse ethical governance systems.

5.1. Promoting Self-Governance of Two-Tier Acting Subjects

The behavior and decision-making entities within large models can be classified into two levels: key groups and corporate entities. The former mainly refers to the developers of large models, while the latter represent economic organizations related to large models that exist in corporate form.

Firstly, strengthening the cultivation and improvement of ethical literacy and professional skills in the key groups of large-scale models is essential. As humans are the source of ethical risks, and the ultimate landing point of ethical governance in large-scale models, the developers of large-scale models form a key group of individuals who possess certain information and spatial advantages. Ethical literacy and professional skills directly impact the original ethical ethics of large-scale models, making them the first line of ethical governance for large-scale models. The ethical self-governance of large-scale model developers can be expanded from enhancing their ethical literacy and professional skills. Public institutions' public-spirited ethical education or promotion serves as a foundational basis, with core curriculum in ethical courses of colleges and universities and supplementary learning or training on the ethical standards and regulations of large-scale models in specific fields. This strengthens the ethical education and learning of large-scale models. The professional skill enhancement requires long-term accumulation and updates that need to be closely followed with technological advancements. In the short term, emphasis should be placed on in-service training, with practical work needs as the orientation, focusing on the integration of theoretical knowledge and practical skills, and closely aligning with technological innovation frontiers to enhance the professional skills of large-scale model technicians. Secondly, in the medium to long term, it is important to leverage higher education as a key pathway, allowing individuals to exercise their autonomy in learning, with employment market demand and individual career development as the ultimate landing points. This cultivates a specialized talent pool.

Secondly, enhance the self-ethical governance of large model enterprises. Large model enterprises mainly include large model development enterprises, large model deployment enterprises, and service provision enterprises. These related enterprises should strengthen their ethical awareness. Ethical considerations should take precedence, integrating concepts such as technology for good, responsible research and technology innovation into corporate culture and practices. Non-formal institutions, such as corporate culture, should play a guiding role in promoting the positive influence of corporate ethical values. The multi-modal, general characteristics, and model-as-a-service (MaaS) supply chain models of large models require the construction of non-formal ethical institutions, such as corporate ethical initiatives or agreements, between enterprises. Meanwhile, companies should strengthen the construction of self-regulatory ethical institutional systems, enhancing the binding power of these systems while broadening their coverage. The self-ethical regulation of large model enterprises exhibits local characteristics, focusing primarily on data security, privacy protection, and user personal information security, and prioritizes the implementation of national ethical regulatory policies and the promotion of corporate development.

Thirdly, innovating self-ethical governance practices, primarily involving strengthening the disclosure of self-ethical governance-related content, and enhancing the application and innovation of large-model ethical autonomy technology or tools. Currently, large-model companies mainly rely on specialized reports such as corporate social responsibility (CSR) reports, ESG reports, and ethical guidelines. It is essential to further extend the disclosure of ethical autonomy-related content from responsible algorithms, responsible artificial intelligence to responsible large models. The technology or tools for large-model ethical review, testing, evaluation, and remediation can help mitigate the occurrence of ethical violations and enhance the safety, reliability, and robustness of large models.

5.2. Leveraging the Role of Two-Level Gatekeepers

The ethical governance role of digital gatekeepers to promote platform neutrality should continually progress. The EU's Digital Markets Act provides a detailed definition of digital gatekeepers, which primarily refer to super platform and large platform companies. Digital gatekeepers can leverage their data, algorithms, and market advantages or powers to acquire certain ethical governance influence. This is achieved through the formation of a strong network effect that results in a consortium-based system centered on platforms, and which contributes to the ethical governance of this system. Therefore, digital gatekeepers play a dual role of self-regulatory gatekeepers and regulatory gatekeepers, bridging between the government and other corporate entities.

From the perspective of platform enterprises' increased obligations and the implementation of social responsibilities, further strengthen the role of secondary digital gatekeepers. 1. Play the role of self-regulating gatekeepers for super platform enterprises and large platform enterprises. Digital gatekeepers establish trust relationships with users on the basis of fulfilling government ethical regulatory policies and assuming corresponding violations, extending to the social trust system. Firstly, establish foundational trust relationships with users, such as clear ownership of user data, data security, and privacy protection, on the basis of current user privacy settings or information collection. Clearly define users' ownership, carryability, and right to be forgotten, etc., and protect user data security and privacy in the process of storing, using, and destroying user data, especially preventing the misuse and illegal trade of user data. Secondly, establish processual trust relationships in platform applications such as default settings, access channels, and user recommendations. For example, some user privacy statements have a certain formality, which exposes the risk of ethical violations and data abuse by transferring data usage rights and actual control rights to users. Processual trust relationships are indispensable. Lastly, establish resultative trust relationships in the provision or output of services and outputs. 2. Play the regulatory gatekeeper role of super platform and large platform enterprises. Enhance the ethical "quasi-regulatory" role of platform enterprises on users regarding deployment areas, input content, and generated content. Leverage their control to take interventions or self-imposed penalties for users violating social ethics rules, regulatory policies, and platform self-made regulatory rules, thereby generating deterrents and cautionary effects within the joint corporate system.

The government must adhere to its ultimate gatekeeper role and accelerate the pace of ethical regulatory system supply and regulatory agency set-up. It should innovate regulatory tools and approaches. The government is the primary supplier of ethical regulatory policies for large-model systems. It should gradually establish a comprehensive ethical regulatory system. To address the challenges posed by the potential ethics violations of large-model systems, the system should enhance its forward-looking, global, flexible, and applicable aspects. Government-affiliated research institutions should strengthen their research on ethical issues related to large-model systems, keeping pace with the forefront of large-model technology development and playing a role in offering advice. It is important to be oriented towards the potential ethics risks, and to proactively deploy the establishment of ethical regulatory systems. Simultaneously, the system should enhance its overall coverage, aiming to encompass the key elements and lifecycle of large-model systems. Large-model systems should be the regulatory objects, key elements the core of regulation, and lifecycle the regulatory scope. It should change the situation where data providers, algorithm providers, and service providers are all separated in regulation. It should establish a comprehensive system of ethical regulatory policies. The flexibility and applicability of ethical regulations also deserve attention. Based on the principle that application scenarios, risk levels, and ethical governance methods should match, on the basis of the "sector-specific regulation" concept, the "scene-specific regulation" concept should be introduced, with priority given to key scenarios and high-risk levels, realizing the flexibility and applicability of ethical regulations. By focusing on key scenarios and high-risk levels, strict legal regulation should be deployed in these areas, clearly defining ethical responsibilities of key elements and entities, setting out clear legal redlines and bottom lines, and enhancing the rigidity of legal constraints. Conversely, for non-priority scenarios and low-risk fields, a soft ethical regulatory system primarily focused on opinions and methods is primarily deployed.

Additionally, the government should set up corresponding large model ethical supervision institutions, and pay attention to the formation and optimization of collaboration and communication mechanisms between different levels of ethical supervision agencies. Considering the differences between unitary, federal, and confederated countries, or unions, a two-tiered ethical supervision system is necessary. A comprehensive large model ethical supervision institution should be established at the central (federal) government, alliance headquarters level, primarily responsible for performing ethical audits, reviews, assessments, investigations, relief, and monitoring of the implementation of ethical supervision systems, among others. Among these, ethical audits are an extension of the current "registration system", reflecting the enhancement of ethical regulatory constraints. Ethical supervision institutions should be established at the local (state) and member country levels, which are administratively subordinate to the aforementioned ethical supervision institutions and have similar functional configurations. Forming or optimizing the operation mechanisms of the two-tiered ethical supervision system, the core is to achieve unified coordination and cooperation and communication, ensuring the enforcement of ethical supervision systems, preventing the occurrence of ethical violations, and handling related ethical violations. Additionally, exploring the establishment of a chief large model ethical governance officer within the supervision agency, who oversees and plans the ethical supervision matters of the institution, and forms a unified and specialized ethical supervision system. Optimizing the large model ethical supervision team, innovating regulatory tools. Large model applications such as generative AI fall within the realm of technology, while ethical supervision is within the field of social sciences. Large model ethical supervision lies at the crossroads of these two disciplines, making it difficult for a single discipline to achieve comprehensive and applicable ethical supervision. Therefore, in the talent pool for the ethical regulation of large models, it is essential to include professionals, including operational technicians and industry experts.

5.3. Global Collaborative Governance Network

The governance of large language models represents a global governance challenge. To address this, global and regional ethical governance cooperation and communication should be strengthened, with global or regional international organizations acting as leading roles, governments playing a significant form of collaboration and communication, and ethical governance innovation as a specific manifestation. At the same time, the public opinion system should play a crucial role in ethical supervision and together form a robust global governance network. The establishment of this network will contribute to the greening of the "deficit" in large model ethical governance, promoting global collaboration and governance.

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

Global international organizations make overall arrangements for the governance of ethics networks, while regional multilateral or bilateral organizations make layouts in accordance with the needs of the local context, primarily through non-mandatory instruments such as declarations, guidelines, and standards. In addition to institutional construction, international organizations should fully leverage their call to action and organizational coordination capabilities, conducting research and exchanges around topics such as the development of large-scale artificial intelligence technology and ethics governance, providing intellectual support for the formulation of ethics governance systems, enhancing cooperation and communication among different countries and regions, and gradually forming a synergistic force for ethics governance. Strengthen government-to-government and regional-to-regional collaboration and exchanges in ethics governance. Leverage the cross-border flow of training data, user data, and algorithms, and international digital trade, among other practical needs, as leverage points. Focus on cooperation contracts and technology collaboration agreements as entry points. Set short-term objectives such as enhancing data security and technology confidentiality. Incorporate ethics governance content related to large models into these contracts and agreements, continuously driving the evolution of ethics governance from bilateral mutual recognition and trust to multilateral mutual recognition and trust, and eventually to global mutual recognition and trust. Innovate ethics governance technology, realizing the governance of technology by technology. Enhance the innovation capabilities of global and multilateral technology organizations, while the foundation of large-scale artificial intelligence technology is the governance of technology. Ethical lapses such as model black boxes and data copyright infringement can be attributed to the current technology that has not yet solved human mysteries or ventured into new fields. Strengthening innovation in large-scale model technology. Collaborating with global, regional, or national technology organizations, focusing on key algorithm, data processing, and privacy computing technologies, to overcome the societal ethics system impacts caused by the uncertainty of large model generation content. Adding an ethical training and optimization component to large model technology, aligning the original and derivative moral and ethical systems with human values. Leveraging the social public opinion system's ethical supervision role. The public is the user of large model applications, the beneficiary and the victim, the largest group of ethical supervision for large model technology. It is an effective supplement to the existing measures such as ethical governance systems, institutions, and technologies, having characteristics such as flexibility, adaptability, timeliness, and wide coverage.

6. Conclusion

The advent of large models signifies a transition from AI technologies with specialized and single-modal properties to those with general applicability and multiplexity. This advancement has greatly facilitated the digital and intelligent transformation of human production and life. However, this development also brings forth ethical misgivings and governance challenges. This paper, drawing from an integrated analysis and review of existing research, utilizes the theories of productivity and production tools to delve deeply into the characteristics of large models and the root causes of their ethical misgivings. The study reveals that large models, as productivity, manifest a technical neutrality, while large models, as production tools, embed human ethics. Due to the dual nature of large models as productivity and production tools, they amalgamate human ethics and machine ethics. This endows large models with an ethical dimension, which differentiates them from traditional productivity and production tools. From a technical and life-cycle perspective, the key elements of large model ethical misgivings are unraveled, revealing that their technical origin is algorithms, the key mediators are data, and the actors are humans.

Addressing the ethical violations inherent in large-scale models, such as model black-box, data copyright infringement, and the difficulty of determining accountability, poses challenges to current ethical governance paradigms employed for single elements and processes, top-down, and result-oriented approaches. Against this backdrop, this study constructs a comprehensive ethical governance framework for large-scale models, focusing on key elements and covering the entire life cycle of the model. By effectively mitigating and regulating the ethical violations of large-scale models, this study aims to promote the effective operation of the new ethical governance framework for large-scale models. To this end, it constructs a two-level behavior system with self-governance, a two-tier gatekeeper, and a global collaborative governance network, forming an ethical governance ecosystem for large-scale models. This accelerates the greening process of the "ethical governance deficit" in large-scale models.

Considering the current research status of large-scale model technology evolution, ethical misconduct characteristics, and related ethical governance processes, future research can be deepened in the following areas. Firstly, research can be conducted from the perspective of large-scale model application scenarios, refining the classification of large-scale model application scenarios, and implementing different ethical governance models for different application scenarios to achieve fine-grained and precise governance of large-scale models. Secondly, research can be conducted from the perspective of large-scale model innovation and iterative updates. By focusing on key ethical issues such as algorithm black box and model black box, research can be conducted on how to achieve algorithmic governance of algorithms and model governance of models from both technical and theoretical perspectives. Thirdly, research can be conducted from the legal perspective, addressing the issues of responsibility attribution, responsibility delineation and punishment measures in large-scale model ethical misconduct, from a systematic perspective.

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

Reference

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

AI Safety Summit 2023: The Bletchley Declaration. (2023, November 1). GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration

Alto, V. (2023). *Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4*. Packt Publishing Ltd.

Boyd, K. L. (2021). Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2), 438:1-438:27. https://doi.org/10.1145/3479582

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (Vol. 95, pp. 11–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_2

CISA and UK NCSC Unveil Joint Guidelines for Secure AI System Development | CISA. (2023, November 26). https://www.cisa.gov/news-events/alerts/2023/11/26/cisa-and-uk-ncsc-unveil-joint-guidelines-secure-ai-system-development

de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial Intelligence Regulation: A framework for governance. *Ethics and Information Technology*, *23*(3), 505–525.

https://doi.org/10.1007/s10676-021-09593-z

Directive—2019/790—EN - dsm—EUR-Lex. (2019). https://eur-lex.europa.eu/eli/dir/2019/790/oj

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* (arXiv:2303.10130). arXiv. https://doi.org/10.48550/arXiv.2303.10130

EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act. (2023). https://artificialintelligenceact.eu/

Ferrara, E. (2024). *Charting the Landscape of Nefarious Uses of Generative Artificial Intelligence for Online Election Interference* (arXiv:2406.01862). arXiv. https://doi.org/10.48550/arXiv.2406.01862

Ford, R. C., & Richardson, W. D. (1994). Ethical decision making: A review of the empirical literature. *Journal of Business Ethics*, *13*(3), 205–221. https://doi.org/10.1007/BF02074820

G20. (2019). *G20 AND Artificial Intelligence*. Center for AI and Digital Policy. https://www.caidp.org/resources/g20/

GPDP. (2023, March 31). *Intelligenza artificiale: Il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori*. https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9870847

GPDP. (2024, January 29). *ChatGPT: Garante privacy, notificato a OpenAI l'atto di contestazione per le violazioni alla normativa privacy*. https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9978020

Gurkan, H., & de Véricourt, F. (2022). Contracting, Pricing, and Data Collection Under the AI

Flywheel Effect. *Management Science*, *68*(12), 8791–8808. https://doi.org/10.1287/mnsc.2022.4333

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. https://doi.org/10.1145/3593013.3594067

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). *A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage*. https://www.authorea.com/doi/full/10.36227/techrxiv.23589741.v1?commit=b1cb46f5b0f749cf5f2f33806f7c124904c14967

Hagendorff, T., & Danks, D. (2023). Ethical and methodological challenges in building morally informed AI systems. *AI and Ethics*, *3*(2), 553–566. https://doi.org/10.1007/s43681-022-00188-y

Hosseini, M., Wieczorek, M., & Gordijn, B. (2022). Ethical Issues in Social Science Research Employing Big Data. *Science and Engineering Ethics*, *28*(3), 29. https://doi.org/10.1007/s11948-022-00380-7

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., & Wang, H. (2024). *Large Language Models for Software Engineering: A Systematic Literature Review* (arXiv:2308.10620). arXiv. https://doi.org/10.48550/arXiv.2308.10620

House, T. W. (2023, October 30). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12), 248:1-248:38. https://doi.org/10.1145/3571730

Karamolegkou, A., Li, J., Zhou, L., & Søgaard, A. (2023). *Copyright Violations and Large Language Models* (arXiv:2310.13771). arXiv. https://doi.org/10.48550/arXiv.2310.13771

King, D. (2024). *Legal & Humble AI: Addressing the Legal, Ethical, and Societal Dilemmas of Generative AI*. Ingene Publishers.

Kowalski, R. (1979). Algorithm = logic + control. *Commun. ACM*, *22*(7), 424–436. https://doi.org/10.1145/359131.359136

Krijger, J., Thuis, T., de Ruiter, M., Ligthart, E., & Broekman, I. (2023). The AI ethics maturity model: A holistic approach to advancing ethical data science in organizations. *AI and Ethics*, *3*(2), 355–367. https://doi.org/10.1007/s43681-022-00228-7

Larsson, S. (2020). On the Governance of Artificial Intelligence through Ethics Guidelines. *Asian Journal of Law and Society*, *7*(3), 437–451. https://doi.org/10.1017/als.2020.19

Li, M., Wan, Y., Zhou, L., & Rao, H. (2024). An enhanced governance measure for deep synthesis applications: Addressing the moderating effect of moral sensitivity through message framing. *Information & Management*, *61*(5), 103982. https://doi.org/10.1016/j.im.2024.103982

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-*

Radiology, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

Lohsse, S., Schulze, R., & Staudenmayer, D. (Eds.). (2019). *Liability for artificial intelligence and the internet of things: Münster Colloquia on EU Law and the Digital Economy IV* (1st edition). Nomos. https://doi.org/10.5771/9783845294797

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., & Buyx, A. (2022). Embedded ethics: A proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*, *23*(1), 6. https://doi.org/10.1186/s12910-022-00746-3

Meeus, M., Shilov, I., Faysse, M., & de Montjoye, Y.-A. (2024). *Copyright Traps for Large Language Models* (arXiv:2402.09363). arXiv. https://doi.org/10.48550/arXiv.2402.09363

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00289-2

Mühlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., & Streit, M. (2014). Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1643–1652. https://doi.org/10.1109/TVCG.2014.2346578

Nakavachara, V., Potipiti, T., & Chaiwat, T. (2024). *Experimenting with Generative AI: Does*

ChatGPT Really Increase Everyone's Productivity? (arXiv:2403.01770). arXiv. https://doi.org/10.48550/arXiv.2403.01770

New and improved content moderation tooling. (2022, August 10). https://openai.com/index/new-and-improved-content-moderation-tooling/

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity* (arXiv:2401.07348). arXiv. https://doi.org/10.48550/arXiv.2401.07348

Pani, B., Crawford, J., & Allen, K.-A. (2024). Can Generative Artificial Intelligence Foster Belongingness, Social Support, and Reduce Loneliness? A Conceptual Analysis. In Z. Lyu (Ed.), *Applications of Generative AI* (pp. 261–276). Springer International Publishing. https://doi.org/10.1007/978-3-031-46238-2_13

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Raiaan, M. A. K., Mukta, Md. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

Schmitt, V., Tesch, J., Lopez, E., Polzehl, T., Burchardt, A., Neumann, K., Mohtaj, S., & Möller, S. (2024). Implications of Regulations on Large Generative AI Models in the Super-Election Year and the Impact on Disinformation. In I. Siegert & K. Choukri (Eds.), *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024* (pp. 28–38). ELRA and ICCL. https://aclanthology.org/2024.legal-1.6

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management (JDM)*, *31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105

Stahl, B. C. (2021). Ethical Issues of AI. In B. C. Stahl (Ed.), *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies* (pp. 35–53). Springer International Publishing. https://doi.org/10.1007/978-3-030-69978-9_4

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence—UNESCO Digital Library*. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Using GPT-4 for content moderation. (2023, August 15). https://openai.com/index/using-gpt-4-for-content-moderation/

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., … Gui, T. (2023). *The Rise and Potential of Large Language Model Based Agents: A Survey* (arXiv:2309.07864). arXiv. https://doi.org/10.48550/arXiv.2309.07864

Xue, L., & Pang, Z. (2022). Ethical governance of artificial intelligence: An integrated analytical framework. *Journal of Digital Economy*, *1*(1), 44–52. https://doi.org/10.1016/j.jdec.2022.08.003

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Yilmaz, E. D., Naumovska, I., & Aggarwal, V. A. (2023). *AI-Driven Labor Substitution: Evidence from Google Translate and ChatGPT* (SSRN Scholarly Paper 4400516). https://doi.org/10.2139/ssrn.4400516

Zhao, J., Khashabi, D., Khot, T., Sabharwal, A., & Chang, K.-W. (2021). *Ethical-Advice Taker: Do Language Models Understand Natural Language Interventions?* (arXiv:2106.01465). arXiv. https://doi.org/10.48550/arXiv.2106.01465

Zhu, Y. (2023, July 10). *Interim Measures for the Administration of Generative Artificial Intelligence Services*. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm

Zorrilla, M., & Yebenes, J. (2022). A reference framework for the implementation of data governance systems for industry 4.0. *Computer Standards & Interfaces*, *81*, 103595.

https://doi.org/10.1016/j.csi.2021.103595

Note: This manuscript was initially released as a preprint on the SSRN on September 12, 2024, representing an early developmental version of the research.