Embedded Way of Responsible Innovation in ChatGPT

Xinyu Chen [1]

[1] *Wuhan University of Technology*
122 Luoshi Road, Wuhan, Hubei, 430070, China

DOI: 10.22178/pos.93-11

LCC Subject Category: B1-5802

Received 25.04.2023
Accepted 28.06.2023
Published online 30.06.2023

Corresponding Author:
chenxinyu7299@163.com

Abstract. In the era of artificial intelligence, ChatGPT, as an advanced language model technology, has the potential for radical innovation. Despite its significant advantages, ChatGPT poses specific potential social and ethical issues. Therefore, we need responsible innovation to mitigate these risks and enable ChatGPT to benefit the global community truly. By embedding responsible innovation throughout the various stages of ChatGPT, we can ensure the practical realisation of public trust in governments and expectations from enterprises, thus achieving compliance and successful implementation. Through such a healthy development approach, we can ensure that ChatGPT positively impacts society and continues to foster its healthy growth.

Keywords: responsible innovation; ChatGPT; ethical issues.

INTRODUCTION

ChatGPT is considered an innovative "next-generation technological revolution". Within just two months of its release, it has garnered over 100 million active users, making it one of the fastest-growing applications in history. We can observe two distinct trends: a comprehensive disruption of the entire industry chain and ecosystem in which it operates and a fundamental transformation of various societal institutions. Given that artificial intelligence has become an irreversible force, the pressing issue we face is how to transition from old paradigms to new ones smoothly. Responsible innovation emerges as a novel concept in the 21st century, aimed at mitigating risks in the research, development, and innovation processes. In this regard, this paper seeks to explore how responsible innovation can be embedded within the entire process of ChatGPT, from its development to its application, providing a reference framework for its sustainable development. The key objective is ensuring that ChatGPT evolves responsibly by addressing ethical considerations and potential biases and minimising harmful consequences. This necessitates a multidimensional approach that encompasses transparency, accountability, and inclusivity. By incorporating responsible innovation principles, ChatGPT can be guided towards a positive and beneficial impact on society while fostering public trust and confidence. The proposed exploration of responsible innovation in the context of ChatGPT acknowledges the significant challenges posed by the rapid advancement of artificial intelligence. It recognises the importance of proactively addressing ethical concerns and societal implications, thereby facilitating the development of a responsible and sustainable AI ecosystem.

RESULTS AND DISCUSSION

ChatGPT's risk environment minutes

As a form of Artificial Intelligence Generated Content (AIGC), Chat GPT has brought "super-simulated" delightful experiences to humans while raising ethical considerations regarding technological advancements [1]. Many countries have adopted a cautious approach towards introducing Chat GPT, with some implementing restrictive regulations. Undeniably, Chat GPT has become the tipping point in the development of artificial intelligence, transitioning from specialised AI to general AI and shifting from niche applications to widespread adoption [2].

The Origin and Development of the Concept of ChatGPT. Before delving into the research on ChatGPT, it is necessary to have a clear understanding and definition of the concept. ChatGPT is an abbreviation for "Chat Generative Pre-

Trained Transformer," which refers to a generative pre-trained transformer model designed explicitly for chat-based interactions [3]. As mentioned earlier, ChatGPT was developed and launched by OpenAI and is a large-scale language model based on the GPT-3.5 architecture. ChatGPT is a machine learning system based on artificial intelligence technology, specifically a large-scale language model implemented in the internet domain using a Transformer architecture [4]. It leverages AI-generated content (AIGC) to generate code, engage in text-based question-answering, and produce digital content. It combines various technical models, including machine learning and neural networks, to train a massive pre-trained language model that learns from human feedback [5]. While ChatGPT can be understood as a software tool, its conceptual definition should transcend the limitations of a chat tool or software enabling chat functionality.

The development of ChatGPT can be traced back to advancements in natural language processing, pre-trained language models, the evolution of the GPT series, as well as the demands and challenges of dialogue systems. Firstly, deep learning and neural network progress have driven significant advancements in natural language processing. Researchers have started utilising large corpora and powerful computational resources to train models capable of generating more accurate and fluent natural language. Secondly, pre-trained language models have gained significant attention in the natural language processing community. By pre-training models on a vast amount of unlabeled data, they can learn rich language representations and semantic understanding, paving the way for more natural and coherent conversations. The evolution of the GPT series, introduced by OpenAI, is another contributing factor. The GPT (Generative Pre-Trained Transformer) models are a series of pre-trained language models based on the Transformer architecture. From GPT1 to GPT3, the models have increased in size and performance, demonstrating exceptional capabilities in various natural language processing tasks and attracting widespread attention and applications. Lastly, the demands and challenges of dialogue systems have played a role. With the proliferation of intelligent assistants and chatbots, there is a growing need for more intelligent and interactive dialogue experiences. However, existing dialogue systems still face challenges in generating high-quality, logically consistent, and naturally interactive responses. In response to these factors, ChatGPT emerged to explore how advanced natural language processing and pre-trained models can build systems that generate coherent and logically consistent dialogues. The focus is optimising training methods, data processing, and user feedback mechanisms to provide better dialogue experiences while ensuring responsible innovation by adhering to ethical, legal, and safety guidelines throughout the system's usage.

Risk Types and Characteristics of ChatGPT. Artificial intelligence (AI) is widely adopted through various algorithmic applications and is in the stage of overcoming challenges to achieve higher levels of development. During this stage, we face two main issues: the difficulties in technological breakthroughs and the incomplete state of ethics, morals, and legal systems. These social factors significantly impact the development of science and technology. Therefore, before evaluating its potential risks, it is necessary to assess the inherent risks of ChatGPT. These risks primarily include technological, industry, ethical, and government risks.

Regarding technological risks, ChatGPT is trained on large datasets from the internet, which may contain biased or controversial content. As a result, ChatGPT may generate responses that reflect or amplify the biases present in the training data. Additionally, ChatGPT cannot self-assess the quality or accuracy of its responses. It may generate answers that sound plausible but are incorrect or absurd without recognising its errors. This could mislead users or provide false information without any evident signs of uncertainty. Like other machine learning models, ChatGPT is susceptible to adversarial attacks. Malicious actors can intentionally input specific phrases or word sequences to manipulate or exploit the model's behaviour, leading to unfavourable or harmful outputs.

Industry risks refer to the impacts on the labour market and the overall economy caused by technological advancements. The birth of new technology is like a double-edged sword. While we desire the convenience it brings, we must also face its harmful effects. Some adverse effects can be anticipated, such as the disappearance of traditional jobs like customer service and telemarketing in the era of ChatGPT.

On the government front, there are a series of risks and challenges. On the one hand, there is widespread capital optimism towards new tech-

nologies. On the other hand, governments must pay closer attention to the societal impacts of ChatGPT. If ChatGPT is widely deployed globally, it will likely conflict with existing urban management systems. However, the development process of ChatGPT is characterised by decentralisation, secrecy, and difficulties in interpretation. Governments cannot implement pre-existing regulations as swiftly as before. When the pace of technological progress surpasses the improvement of government governance, technology can influence the political system and threaten government authority.

Ethical risks mainly involve issues of responsibility attribution and allocation. ChatGPT is a system based on automated algorithms. It may make incorrect judgments or provide inaccurate information in certain situations. In such cases, the determination and allocation of responsibility are not decided by ChatGPT itself but by the algorithms written by the developers. Therefore, the attribution and assignment of responsibility become highly complex, potentially leading to dilemmas where fault cannot be determined. More urgently, there is no consensus on allocating certain ethical risks at the societal level.

Pre-regulation is advocated to mitigate the risks and ensure responsible innovation in the accelerating era of AI. Responsible innovation involves pre-emptively preventing and controlling product risks during the design stage. For ChatGPT, responsible innovation needs to be embedded throughout all stages to maximise its service to society and enhance social well-being rather than the opposite. Such measures will effectively manage the risks of the AI era.

Responsible Innovation Provides Legitimacy for ChatGPT Implementation

As ChatGPT is widely adopted, the public's expectations for artificial intelligence continue to rise. However, concerns about its negative impacts have also emerged, leading to questions about the responsibility attributed to this innovative technology. "The most influential policy impact in this trend is the concept of Responsible Research and Innovation (RRI) proposed at the European Union level" [6]. Currently, there is no unified and precise definition of responsible innovation. Still, Schomberg's report provides a comprehensive perspective: "Responsible research and innovation is a transparent interactive process in which societal actors and innovators mutually respond, fully considering the (ethical) acceptability, sustainability, and societal desirability of the innovation process and its market products, to embed technological developments into our society appropriately" [7]. In the European Commission's Horizon 2020 framework program, it is defined as an approach that "anticipates and assesses the potential implications and societal expectations of research and innovation, aiming to help design inclusive and sustainable research and innovation" [8]. Its characteristics involve incorporating more elements into the responsibility system, considering human rights more, and pursuing the greening and democratisation of innovation outcomes to manage (technological) innovation practices and strive to maximise the benefits of innovation for society [9]. From a developmental perspective, it is crucial to integrate ethical factors into the entire innovation process and determine whether researchers have the necessary moral knowledge to meet societal and public requirements, ultimately enabling the public to make responsible choices with informed consent. In the long run, "governing issues of technological innovation design and analysis urgently need to be introduced into a more systematic innovation paradigm framework while promoting industrial innovation, creating growth opportunities, and guiding social transformation" [10].

Embedding Responsible Innovation into the Path of ChatGPT. The future direction of ChatGPT cannot be solely determined based on its inherent characteristics. Companies like OpenAI use this new technology to establish a technological barrier and maintain a distance from traditional companies. However, scientific research and development have also entered a relatively unfamiliar exploration domain. As a technology with a complex architecture, ChatGPT bears complex ethical responsibilities. Therefore, the following requirements must be met during the design and development stages. Firstly, the research and development personnel involved in the design and development must be proactive and responsible for the products they design. This responsibility should be shared among the entire team involved in the design and development stages. Secondly, responsibility standards should be integrated into the technical systems of designers and R&D personnel at all levels.

When responsible innovation is integrated into the technical aspect of ChatGPT, several key con-

siderations should be addressed. Firstly, it is essential to define the usage scenarios and target user groups of the chatbot. This helps to consider relevant responsible innovation issues during the development process and ensures that specific user needs are met. Secondly, it is crucial to train the ChatGPT model using diverse datasets. This means collecting user data from different groups and cultural backgrounds to avoid biases and discrimination. Thirdly, introducing transparency and interpretability mechanisms helps to increase user trust in ChatGPT. Users want to know how the chatbot generates answers and makes decisions. By explaining the model's workings, the data used, and the algorithms employed, more understandable answers can be provided to users. Fourthly, protecting user privacy is an essential aspect of responsible innovation. Measures should be taken to safeguard user's personal information and privacy when using ChatGPT. Fifthly, establishing a robust review mechanism is critical to ensuring the responsible use of ChatGPT. Manual review can help identify and filter out inappropriate content, ensuring that responses align with ethical and legal requirements.

Additionally, establishing a mechanism for handling complaints and addressing user concerns enables timely detection and resolution of potential risks and violations. Lastly, promoting the responsible use of chatbots is paramount. Providing user guidelines and educational training helps users understand how to interact with ChatGPT, and feedback mechanisms can be implemented to collect user opinions and suggestions. At the same time, advocating for ethical use and actively disseminating relevant information can guide users in maintaining appropriate behaviour and expectations when interacting with chatbots.

The Four Dimensions of Responsible Innovation Implementation. Applying the framework of responsible innovation can effectively regulate the development of ChatGPT, reducing future uncertainties and risks. Two prerequisites are necessary to achieve accountable innovation. Firstly, "responsible innovation" is a multidisciplinary concept that requires in-depth discussions among researchers, humanities and social sciences scholars, governments, and businesses, among other stakeholders, regarding the direction of technological development. Secondly, proactive research should be conducted at the early stages of technology development to provide more alternative options for the technology, facilitating its improvement. British scholar Owen pointed out that "responsible innovation, as a technological management approach, should have the basic cognitive characteristics and consists of four dimensions: anticipation, reflection, inclusiveness, and responsiveness" [11]. Based on these four dimensions, ChatGPT can be analysed regarding how they will play a role.

In the dimension of anticipation, designers should continuously consider their design goals to ensure clarity and avoid potential social harm. Conversely, if they assume known and unknown risks, the risks associated with new technologies will significantly reduce. In reflection, as this approach is a developing research method, it requires third parties to objectively and professionally evaluate the researchers' research objectives, motivations, and decisions. Such evaluations act as a mirror for researchers, enabling them to gain a clear understanding of whether their goals comply with regulations. In the dimension of inclusiveness, in addition to seeking expert advice, extensive public comments and consultations should be conducted to bridge the understanding gap among relevant stakeholders. This ensures the broadness and universality of the research outcomes and facilitates their application in social contexts, effectively addressing societal concerns. In the responsiveness dimension, responsible innovation requires developers to maintain sensitivity and promptly respond to customer feedback, suggestions, and risks, promptly adjusting technical solutions and research priorities. This feedback process involves a two-way interaction where companies or developers provide guidance and input to government agencies when formulating policies to enhance the quality of policy-making.

In the practice of responsible innovation, corresponding measures can be proposed from these perspectives to address better the various challenges that may arise during the development of ChatGPT. It is essential to integrate the needs of multiple stakeholders comprehensively. Only through such decentralised moral distribution can responsible innovation achieve its goals.

Responsible Innovation Mechanisms for ChatGPT

The implementation of responsible innovation involves three key issues. First, it requires the involvement of relevant professionals and stake-

holders, along with appropriate adjustments to ensure information transparency and rights constraints. Second, throughout the entire process of technological innovation, it is essential to leverage the expertise and engagement of experts and various societal actors to overcome the limitations of traditional "narrow innovation responsibility." Third, it involves using policy and legal enforcement measures to standardise the development process of science and technology. "In the field of technology and innovation politics, authoritative decisions made by law enforcement, legal, and administrative departments through orders, prohibitions, and the allocation of functions are the normative guidelines that must be followed" [12].

Therefore, several aspects should be considered regarding responsible innovation mechanisms for ChatGPT. Firstly, establishing monitoring and filtering mechanisms to identify and prevent inappropriate or harmful content is crucial. This can be achieved through content filtering algorithms, human moderation, and user feedback mechanisms. It ensures that the generated responses by ChatGPT align with ethical, legal, and societal norms, avoiding disseminating misinformation, discriminatory speech, or other objectionable content. Secondly, providing transparency and interpretability is critical to achieving responsible innovation. Users need to understand how the ChatGPT model works and how it generates responses. This can be accomplished through technical documentation, public disclosure of model architecture, and explanatory features in the user interface. Transparency and interpretability help users comprehend the decision-making process of ChatGPT and reduce potential misunderstandings or mistrust. Thirdly, encouraging user engagement and feedback is a meaningful way to promote responsible innovation. Users can participate in the improvement process of ChatGPT by providing ratings, reporting inappropriate behaviour, or suggesting enhancements. Such user involvement helps identify potential issues and biases, prompting the technical team to improve and optimise the functionality and performance of ChatGPT continuously. Lastly, continuous monitoring and updates are crucial to ensure ongoing responsible innovation. The technical team needs to stay abreast of the latest ethical guidelines and societal expectations, incorporating them into the development and updating process of ChatGPT. This may involve regular reviews of model performance and impact, necessary corrections and improvements, and ensuring that ChatGPT adapts to the evolving social and technological landscape.

Through these responsible innovation mechanisms, ChatGPT can provide high-quality services while proactively addressing ethical, legal, and societal risks, ultimately achieving the goal of responsible innovation.

CONCLUSIONS

The launch of ChatGPT will trigger a significant social transformation. However, if we adhere to past innovation models, new technologies represented by ChatGPT are likely to cause social panic and potential risks. This is because it involves a unique collaboration mechanism and value orientation. Therefore, as a new paradigm, responsible innovation should be integrated throughout the entire process, from production to consumption, and the allocation and matching of responsibilities become inherent in purposeful innovation. This requires us to incorporate the concept of responsible innovation throughout the entire process of ChatGPT and leverage the power of government and society to guide and supervise its development. Companies should also assume corresponding corporate responsibilities. This article provides suggestions on how technology can participate in the responsibility-sharing mechanism. Only with awareness and collaboration from all parties can ChatGPT have a bright future, ensuring that new technologies provide benefits to society rather than risks.

Acknowledgements

This study was supported by “The Fundamental Research Funds for the Central Universities” (grant number: 2023VB067).

REFERENCES

1. Wang, J., & Cao, H. (2023). Research on the Communication Characteristics, Logic and Paradigm of ChatGPT. *Journal of Shenzhen University (Humanities & Social Sciences), 40*(2), 144-152 (in Chinese).

2. Zhong, X., & Fang, X.（2023). Governance of ChatGPT: Challenges and Countermeasures. *Media Observer, 471*(3), 25–35 (in Chinese).

3. Klie L. (2023). OpenAI Introduces ChatGPT, a New Al Chatbot Model. *CRM Magazine, 27*(1):10-11.

4. van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature, 614*(7947), 224–226. doi: 10.1038/d41586-023-00288-7

5. Zhang, X. (2023). Research on the Development of Digital Economy Based on New Generation Artificial Intelligence Technology (ChatGPT). *Journal of Chang'an University (Social Science Edition), 1*(02), 1–10 (in Chinese).

6. Liao, M. (2019). Responsible Research and Innovation in the Intellectual History of European Policy. *Studies in Science of Science, 37*(7), 1212–1219 (in Chinese).

7. von Schomberg, R. (2012). Prospects for technology assessment in a framework of responsible research and innovation. *Technikfolgen Abschätzen Lehren*, 39–61. doi: 10.1007/978-3-531-93468-6_2

8. European Commission. (2016, November 17). Responsible Research & Innovation. Retrieved from https://commission.europa.eu/funding-tenders_en

9. Liu, Z. (2015). A Review of Responsible Innovation Research: Background, Current Situation, and Trends. *Science & Technology Progress and Policy, 32*(11), 155-160 (in Chinese).

10. Brundage, M. (2016). Artificial Intelligence and Responsible Innovation. *Synthese Library*, 543–554. doi: 10.1007/978-3-319-26485-1_32

11. Stillgoe, J., Owen, R., & Macnaghten, P. (2020). Developing a Framework for Responsible Innovation. In A. Maynard, J. Stilgoe, *The Ethics of Nanotechnology, Geoengineering and Clean Energy* (pp. 1568–1580). doi: 10.4324/9781003075028

12. Grunwald, A. (2018). *Handbook of Technoethics.* Beijing: Social Sciences Academic Press.