Sharing Accountability of Versatile AI Systems: The Role of Developers and Practitioners

Changhyun Lee[a]**, Hun Yeong Kwon**[b]**,** and **Kyung Jin Cha**[a]
[a]Hanyang University, 222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea
[b]Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea
newdlckdgus@hanyang.ac.kr, khy0@korea.ac.kr, kjcha7@hanynag.ac.kr

Abstract

AI systems pose both opportunities and threats in various industries. To harness these opportunities and mitigate risks, accountability is crucial. Traditionally, developers bear the responsibility for auditing and modifying algorithms. However, in the evolving landscape of versatile AI, developers may lack contextual understanding across diverse fields. This paper proposes a theoretical framework that distributes accountability to developers and practitioners according to their capabilities. This framework enhances systemic comprehension of shared roles, empowering both groups to collaboratively avert potential adverse impacts.

Introduction

As the influence of artificial intelligence (AI) on daily life continues to grow, concerns regarding its potential adverse effects have become more pronounced (Arrieta et al., 2020). Consequently, the need for accountability to mitigate these potential adverse effects has been a focal point in numerous studies (Diakopoulos, 2016). This responsibility is commonly attributed to AI developers, who possess the capability and authorization to modify the model (Arrieta et al., 2020; Shin, 2021).

However, the recent advent of versatile AIs, an AI applicable for various purposes such as large language models (LLMs), has cast doubt on the current perspective of accountability. In contrast to earlier narrow AIs, designed for specific purposes, developers of versatile AIs face the challenge of not being able to fully anticipate the dynamic contexts across diverse fields and industries where the AI may be susceptible to abuse or misuse (Sallam et al., 2023).

In this instance, accountability should be jointly shouldered by practitioners who possess a profound understanding of the specific context in which versatile AI is implemented. However, practitioners frequently lack the capability and authorization necessary to audit or modify the algorithm (Dwivedi et al., 2023). Therefore, there is a need to reassess the distribution of accountability for versatile AI, considering the capabilities of both developers and practitioners (Burr & Leslie, 2023). Consequently, the objective of this study is to delve into the concept of accountability and examine the capabilities of developers and practitioners to delineate their responsibility for the potential adverse effects of versatile AI.

This study suggests the responsibility of developers and practitioners on versatile AI, considering their capability based on the theories of digital transformation and accountability. We review the literature on digital transformation of versatile AI and accountability to categorize the accountability and suggest the developer's roles and practitioner's roles for each category. The paper is concluded with the brief conclusion on the contributions and limitations of this study.

Literature Review

Digital Transformation of Versatile AI

On the broad spectrum, AI is categorized into narrow AI and general AI. Narrow AI pertains to an AI designed to execute specific tasks, while general AI refers to an AI endowed with abilities comparable to or exceeding those of humans (Schlegel & Uenal, 2021; Gutierrez et al., 2023). Currently, only narrow AI is in practical implementation, with scholars aspiring to achieve general AI in the long term.

Nevertheless, the recent advancements in LLMs have introduced versatile functionalities, marking the initiation of a transition from narrow AI to general AI. An illustrative example is ChatGPT, a chatbot based on LLMs extensively trained using the Generative Pre-training Transformer (GPT) architecture. Scholars and practitioners across various domains, including education, programming, and communication, widely study ChatGPT (Dwivedi et al., 2023). Although ChatGPT is classified as a narrow AI, capable only of generating natural language based on given requests, its versatility allows it to be applied across diverse fields (Vemprala et al., 2023).

The versatility of such AIs holds the potential to introduce new value propositions and facilitate digital transformation across diverse fields. Digital transformation, broadly recognized as the seamless integration of digital technologies and innovative business models (Vial, 2021), diverges from mere digitization or digitalization by placing significant emphasis on leveraging digital technology to create and provide additional value (Tabrizi et al., 2019). From this standpoint, digital technology is acknowledged as an enabler of digital transformation, aiding in the realization of proposed value propositions. Simultaneously, the business model is identified as the driver, serving as the operational framework that facilitates the delivery of envisioned value (Tekic & Koroteev, 2019). In this context, versatile AIs can facilitate the digital transformation of various domains by serving as an enabler for new business models that were previously unattainable (Dwivedi et al., 2023).

However, the versatility of AI possesses a dual nature, as humans utilizing the model can potentially abuse or misuse the technology. For example, an increasing number of students resort to cheating on assignments with ChatGPT, prompting arguments for the prohibition of its use in education (Johnson, 2023). While the abuse and misuse of AI were already critical issues, this problem becomes more pronounced in the context of versatile AI. This is because controlling its abuse and misuse, which may occur simultaneously in multiple contexts, becomes more challenging (Dwivedi et al., 2023).

Accountability

In the context of AI, accountability refers to the human responsibility of justifying ensuring that AI outputs align with the common good (Diakopoulos, 2016; Mittelstadt et al., 2019). Given the potential for abuse or misuse of AIs across different applications, scholars have concentrated on whether AI is designed to be mitigated when potential adverse effects arise. For instance, Shin (2021) proposed that AI should be developed to facilitate human auditing and control of the algorithm's behavior in a timely manner. Similarly, Novelli et al. (2023) suggested that the algorithm should be designed to generate outputs in alignment with ethical standards (i.e., compliance), ensure the proper recording of the agent's conduct for justification (i.e., report), provide evidence for humans to assess the agent's conduct (i.e., oversight), and determine the consequences the agent must bear (i.e., enforcement).

To regulate the abuse or misuse of algorithms, accountability includes adhering following two primary dimensions: controllability and openness. Controllability pertains to the aspect of human oversight and intervention in auditing and modifying AI system configurations to mitigate irrational or socially undesirable outcomes (Shin, 2021). This emphasis on controllability is particularly pertinent in facilitating direct human influence to prevent AI systems from generating outputs that contradict societal interests (Lee & Cha, 2023).

Openness encompasses the ethical obligation of individuals to offer clear and comprehensible explanations of the rationale behind their decisions to all pertinent stakeholders, irrespective of their level of expertise (London, 2019). The emphasis on openness plays a pivotal role in mitigating adverse consequences arising from information asymmetry, thereby empowering stakeholders to take proactive measures to safeguard their interests (Chiu et al., 2009; Arrieta et al., 2020).

Meanwhile, some perspectives regard accountability as a social commitment aimed at preventing the abuse or misuse of AI (Novelli et al., 2023). In this context, accountability is also associated with competence and benevolence (McKnight et al., 2011). Essentially, for an individual to be accountable for AI, they must possess the capability and goodwill to actively control its adverse effects or communicate them to others (Ryan, 2020).

Theoretical Development

In the context of versatile AI, this study suggests that accountability should be systematically allocated to developers and practitioners based on their capabilities. Developers possess the capability to directly audit and modify the algorithm, but they often lack understanding of the contexts where the AI is applied. Conversely, practitioners have a nuanced understanding of the domain context, but they often lack the techniques to directly modify the model (Shin, 2021; Dwivedi et al., 2023). Therefore, this study delves into the features of digital transformation and accountability to identify the boundaries of developers' and practitioners' accountability. Figure 1 summarizes the discussion on the distribution of accountability.

Enabler Features

Enabler - Controllability

In the context of digital transformation, enabler refers to the model's capacity to facilitate functions that were previously deemed impossible (Tekic & Koroteev, 2019). Features such as performance and explainability can be regarded as enabler features, where high performance enables sophisticated predictions, and high explainability allows human practitioners to reference the decision basis during decision-making (London, 2019). Therefore, the controllability of enabler features pertains to the responsibility of enhancing the algorithm's performance and explainability to minimize adverse effects resulting from malfunctions (Shin, 2021). From this perspective, the controllability of enabler features should be attributed to developers who possess the capability and authorization to modify the algorithm, rather than practitioners, as practitioners often lack the capabilities to audit or modify it (Dwivedi et al., 2023).

Enabler - Openness

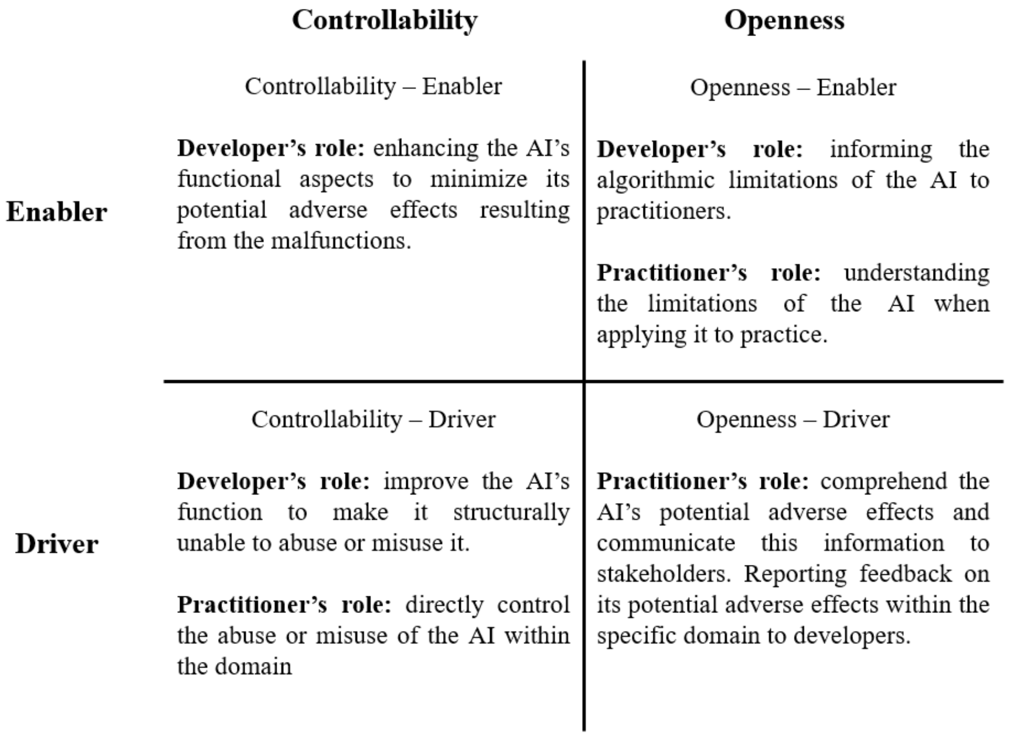
The openness of the enabler, on the other hand, entails the responsibility to comprehend the potential adverse effects of versatile AI and communicate them to stakeholders (Chiu et al., 2009). In this context, developers should apprise practitioners of the performance limitations inherent in the algorithm they have created. Simultaneously, practitioners bear the responsibility to understand the algorithm's limitations based on the developer's explanation and relay this information to stakeholders within the domain. For instance, the developers of ChatGPT should inform educators that ChatGPT may generate inaccurate information, and educators, in turn, should inform students that relying heavily on ChatGPT during learning may lead them astray.

Driver Features

Driver - Controllability

In the context of digital transformation, a driver refers to human planning to leverage technology for new value propositions (Tekic & Koroteev, 2019). Since technologies can be either well-used or abused based on the user's intention (Blauth et al., 2022), malicious user intentions can be identified as a representative driver feature that accountability should scrutinize. In this case, the controllability of the driver pertains to the responsibility of directly controlling the behavior of malicious technology use. Similar to considerations of legal compliance, the controllability of the driver can encompass governmental regulations and social norms (Griffith, 2015). The government should establish a legal mechanism to regulate technology abuse, and social members, including practitioners, should refrain from exploiting technology. Developers also need to design AI to systematically prevent abuse. For example, the government could enact a legal policy to address the creation of fake reviews using ChatGPT, platform managers (practitioners) could directly regulate complementor's advertisements that utilize fake reviews generated by ChatGPT, and developers could enhance their algorithms to make it challenging for ChatGPT to respond to requests for generating fake reviews.

Driver - Openness



	Controllability	**Openness**
Enabler	Controllability – Enabler **Developer’s role:** enhancing the AI’s functional aspects to minimize its potential adverse effects resulting from the malfunctions.	Openness – Enabler **Developer’s role:** informing the algorithmic limitations of the AI to practitioners. **Practitioner’s role:** understanding the limitations of the AI when applying it to practice.
Driver	Controllability – Driver **Developer’s role:** improve the AI’s function to make it structurally unable to abuse or misuse it. **Practitioner’s role:** directly control the abuse or misuse of the AI within the domain	Openness – Driver **Practitioner’s role:** comprehend the AI’s potential adverse effects and communicate this information to stakeholders. Reporting feedback on its potential adverse effects within the specific domain to developers.

Figure 1. The Distribution of Accountability

The openness of the driver, on the other hand, involves the responsibility to comprehend the adverse effects of AI abuse and communicate this information to stakeholders (Lee & Cha, 2023). Given that developers lack an understanding of AI's potential adverse effects within specific contexts, the openness of the driver is primarily ascribed to practitioners rather than developers. Furthermore, as developers require feedback from practitioners to enhance the algorithm systematically and prevent its abuse, the openness of the driver encompasses the practitioner's responsibility to report potential adverse effects of the algorithm to developers.

Discussion and Conclusion

In the era of versatile AI, the ongoing development of algorithms with universal applicability across diverse fields and industries is anticipated. The allocation of accountability for directly controlling and comprehending the potential adverse effects of versatile AI is pivotal in establishing the principles of ethical AI and responsible AI. Through an examination of the capabilities and limitations of both developers and practitioners, this study contributes to a systemic approach to understanding their responsibilities in the realm of versatile AI.

This study offers implications for both academic and practical sectors. Primarily, it critiques the previous approach to accountability, which predominantly centers on the responsibility of developers, who have the capability and authorization to modify the algorithm. Given the nature of versatile AI, where developers may not entirely foresee adverse scenarios of AI abuse or misuse across various domains, this study acknowledges that practitioners, with their understanding of the domain context, bear responsibility for preventing abuse and misuse within that specific context. This approach signifies a paradigm shift in the perspective of accountability, not solely concentrating on the role of developers but also emphasizing the role of practitioners.

Secondly, it unpacks and integrates theories of digital transformation and accountability to elucidate the roles of developers and practitioners concerning accountable versatile AIs. This clarification is poised to aid in the future formulation of laws and policies for versatile AIs by delineating the boundaries of responsibility for each actor. This systemic approach expands the understandings on accountability and enables scholars not to assign responsibility without the bounds of the actor's capabilities.

Finally, this study raises doubts on full automation. Suppose there is a truck in front and a motorcycle on both sides of the road, and heavy cargo carried by the truck is poured out. If the car stops or runs as it is, it will be hit by the cargo and endanger the driver. If a car turns, it can get hit by a motorcycle and crash the motorcycle driver. If the driver is a human, no matter what the driver does, it can be considered a reflex action driven by the survival instinct, so there is little reason to be ethically criticized. However, if the driver is AI, the crux of this dilemma lies in the fact that developers must decide in advance whom to put at risk when building AI algorithms. However, when we consider the accountability of the practitioner, the responsibility for this problem shifts to the driver themselves, who placed themselves in a precarious situation by relinquishing control in a hazardous scenario. In this case, the developer's responsibility is to create a high-performance AI system that minimizes exposure to perilous situations as much as possible. However, drivers have a responsibility to evade themselves from risky situation by considering the traffic context they face. As such, this study introduces a new paradigm for ethical AI and underscores that the accountability of the practitioner should not be disregarded when considering AI accountability in the future.

Despite these implications, this study includes several limitations. Firstly, as a literature review, this study did not provide quantitative evidence for the framework. Therefore, future studies are recommended to conduct systemic research by employing the framework. Secondly, there are potential actors they are responsible to the potential adverse effect of versatile AI. Therefore, future studies may expand our framework, considering the public sector or other third parties.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58,* 82-115.

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access, 10,* 77110-77122.

Burr, C., & Leslie, D. (2023). Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics, 3*(1), 73-98.

Chiu, C. M., Lin, H. Y., Sun, S. Y., & Hsu, M. H. (2009). Understanding customers' loyalty intentions towards online shopping: an integration of technology acceptance model and fairness theory. *Behaviour & Information Technology, 28*(4), 347-360.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56-62.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71,* 102642.

Griffith, S. J. (2015). Corporate governance in an era of compliance. Wm. & Mary L. Rev., 57, 2075.

Gutierrez, C. I., Aguirre, A., Uuk, R., Boine, C. C., & Franklin, M. (2023). A proposal for a definition of general purpose artificial intelligence systems. *Digital Society, 2*(3), 36.

Johnson, A. (2023). ChatGPT In Schools: Here's Where It's Banned—And How It Could Potentially Help Students. Forbes, Retrieved in: https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/?sh=4f3c93c6e2c6

Lee, C., & Cha, K. (2023). FAT-CAT—Explainability and augmentation for an AI system: A case study on AI recruitment-system adoption. *International Journal of Human-Computer Studies, 171,* 102976.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report, 49*(1), 15-21.

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology An investigation of its components and measures. *ACM Transactions on management information systems (TMIS), 2*(2), 1-25.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the conference on fairness, accountability, and transparency*. 279-288.

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY*, 1-12.

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749-2767.

Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare, 11*(6), 887-906.

Schlegel, D., & Uenal, Y. (2021). A Perceived Risk Perspective on Narrow Artificial Intelligence. In PACIS (p. 44).

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146,* 102551.

Shin, D., Zhong, B., & Biocca, F. A. (2020). Beyond user experience: What constitutes algorithmic experiences?. *International Journal of Information Management, 52*, 102061.

Tabrizi, B., Lam, E., Girard, K., & Irvin, V. (2019). Digital transformation is not about technology. *Harvard business review, 13*(March), 1-6.

Tekic, Z., & Koroteev, D. (2019). From disruptively digital to proudly analog: A holistic typology of digital transformation strategies. *Business Horizons, 62*(6), 683-693.

Vial, G. (2021). Understanding digital transformation: A review and a research agenda. *Managing Digital Transformation, 28*(2), 13-66.

Vemprala, S., Bonatti, R., Bucker, A., & Kapoor, A. (2023). Chatgpt for robotics: Design principles and model abilities. Microsoft Auton. *Syst. Robot. Res, 2*, 20.