Taxonomy of Generative AI Applications for Risk Assessment

Hiroshi Tanaka
Fujitsu Limited
Kawasaki, Japan
htnk@fujitsu.com

Masaru Ide
Fujitsu Limited
Kawasaki, Japan
masaru.ide@fujitsu.com

Jun Yajima
Fujitsu Limited
Kawasaki, Japan
jyajima@fujitsu.com

Sachiko Onodera
Fujitsu Limited
Kawasaki, Japan
sachiko@fujitsu.com

Kazuki Munakata
Fujitsu Limited
Kawasaki, Japan
munakata.kazuki@fujitsu.com

Nobukazu Yoshioka
Waseda University
Tokyo, Japan
nobukazuy@acm.org

ABSTRACT

The superior functionality and versatility of generative AI have raised expectations for the improvement of human society and concerns about the ethical and social risks associated with the use of generative AI. Many previous studies have presented risk issues as concerns associated with the use of generative AI, but since most of these concerns are from the user's perspective, they are difficult to lead to specific countermeasures. In this study, the risk issues presented by the previous studies were broken down into more detailed elements, and risk factors and impacts were identified. In this way, we presented information that leads to countermeasure proposals for generative AI risks.

CCS CONCEPTS

• **General and reference→Evaluation;** *Surveys and overviews*;
• **Human-centered computing→***HCI theory, concepts and models*;
• **Social and professional topics→***Computing / technology policy*.

KEYWORDS

language models, responsible innovation, technology risks, responsible AI, risk assessment

ACM Reference format:

Hiroshi Tanaka, Masaru Ide, Jun Yajima, Sachiko Onodera, Kazuki Munakata and Nobukazu Yoshioka. 2024. Taxonomy of Generative AI Applications for Risk Assessment. In *Proceedings of 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN'24). Lisbon, Portugal, 2 pages.*

1 Introduction

Because generative AI, based on highly accurate fundamental models, can be easily utilized by ordinary users with superior functionality not available in conventional AI, there are concerns about the ethical and social risks associated with its use. To address these concerns, studies have classified the impact of generative AI risks into risk domain classes [1][2], and technical documents have been created to highlight safety issues [3]. These studies have been referenced in government AI strategy documents and incorporated into national strategies [4][5].

The primary purpose of a risk study is to develop risk countermeasures. However, the risk issues identified in previous studies are presented with various levels of description (risk factors, risk impacts, etc.), making it challenging to derive specific risk countermeasures. Therefore, we decomposed risk into factors and impact, classifying each into risk domain classes. This approach enables users to present key considerations when using generative AI and corresponding countermeasures in an easy-to-understand manner.

In this paper, we first present 20 risk issues in Section 2, consolidating the generative AI risks described in previous studies, followed by the risk domain classes further subdivided from the six classes outlined in the papers [1][2]. In Section 3, we present the decomposition results of risk into factors and impacts, clarifying the relationship with risk countermeasures. Finally, we provide a conclusion and discuss future perspectives.

2 Risk issues and risk domain classes

Numerous studies [1-5] have identified risks associated with generative AI, but these risks are not identical across the studies (e.g., 21 risks classified into 6 categories [1][2], 12 risks identified [3], etc.). We consolidated and organized these into 20 risk issues (Table 1) and created detailed risk classes [1][2] (Table 2).

These risk issues are outlined from the user's perspective, with the factors and effects of risk blended together. This makes it difficult to clearly understand the necessary steps for risk mitigation and the specific improvements that should be aimed for in risk measures.

3 Decomposing risks into factors and impacts

According to the safety standard ISO/IEC Guide 51 [6], risk can be separately modeled as hazard and impact. This model posits that improper system behavior (hazard) is a factor that increases

CAIN 2024, April 14–15, 2024, Lisbon, Portugal
 ISBN 979-8-4007-0591-5/24/04.
https://doi.org/10.1145/3644815.3644977

the likelihood of damage, and defines risk as the expected value of damage when a hazard occurs. Following this concept, Figure 1 simplifies the process of AI risk occurrence. To mitigate the hazard of generative AI risk, it is essential to distinctly separate hazard and impact, which the risk issue represents.

Based on this concept, we have divided the risk issue into hazard and impact (Table 3). This enables us to associate risk reduction measures with improvement effects on impacts, while concentrating on the hazard of the risk issue.

4 Discussion and Conclusion

To mitigate risk, we can: 1) remove the risk source, 2) avoid the hazard, and 3) manage the impact (Figure 1). Table 3 helps with measures 2) and 3), but measure 1) needs a risk analysis considering the AI system's configuration. Our next step is to apply a framework for analyzing risk occurrence and its impact on AI systems, such as AIEIA [7].

REFERENCES

[1] L. Weidinger, et.al. "Ethical and social risks of harm from Language Models," arXiv:2112.04359 [cs] (Dec. 2021).

[2] L. Weidinger, et.al. "Taxonomy of Risks posed by Language Models," Proc. of FAccT '22, pp.214- 229, DOI: 10.1145/3531146.3533088 (June 2022).

[3] OpenAI, "GPT-4 System Card," (Mar. 2023). https://cdn.openai.com/papers/gpt-4-system-card.pdf

[4] A. KATIRAI, K. Ide, A. Kishimoto, "Overview of the Discussion Points on Ethical, Legal, and Social Issues (ELSI) of Generative AI (Generative AI) : March 2023 Edition", Osaka Univ. ELSI NOTE. 2023,26, pp.1-37, DOI : 10.18910/90926 (March 2023). (In Japanese)

[5] CRDS JST, "New Trends in Artificial Intelligence Research 2 - Impact of Fundamental Models and Generative AI," Strategic Proposal/Report CRDS-FY2023-RR-02, (July 2023). https://www.jst.go.jp/crds/report/CRDS-FY2023-RR-02.html. (In Japanese)

[6] ISO/IEC Guide 51:2014 Safety aspects - Guidelines for their inclusion in standards, https://www.iso.org/standard/53940.html

[7] I. Nitta, K. Ohashi, S. Shiga and S. Onodera, "AI Ethics Impact Assessment based on Requirement Engineering," Proc. of 30th Intl. Requirements Engineering Conference Workshops (REW), Melbourne, Australia, pp.152-161, DOI: 10.1109/REW56159.2022.00037 (Aug. 2022).

Table 1: Risk Issues

	Risk issue		Risk issue
1	Hallucination	11	Economic impacts
2	Potential for risky emergent behaviors	12	Acceleration
3	Harmful content	13	Environmental and financial cost
4	Harms of representation, allocation, and quality of service	14	Spreading misinformation
5	Disinformation and influence operations	15	Increasing sophistication and ease of crime
6	Overreliance	16	Proliferation of conventional and unconventional weapons
7	Privacy	17	Illegal surveillance and censorship
8	Copyright infringement	18	Lack of transparency of training data
9	Exploitation of workers during model creation	19	Interactions with other systems
10	Cybersecurity	20	No rights (copyrights or patents) for AI creations

Table 2: Risk Domain Classes

class	Major class of risk	subclass	Subclass of risk
1	Discrimination, Hate speech and Exclusion	1-1	Toxic Content Generation
		1-2	Social Effects of Unfair Discrimination
2	Information Hazards	2-1	Information Leakage
		2-2	Right Infringement
3	Misinformation Harms	3-1	Misinformation Output
		3-2	Biased Information Output
4	Malicious Uses	4-1	Intentional Harmful Content Generation
		4-2	Cybersecurity Decline
5	Human-Computer Interaction Harms		
6	Environmental and Socioeconomic Harms	6-1	Deterioration of Social Environment
		6-2	Deterioration of Information Environment
		6-3	Economic Damage

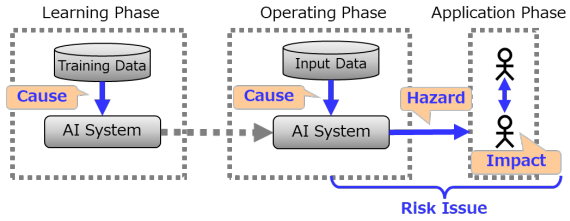

Figure 1: Model of Risk Occurrence

Table 3: Decomposition of Risk Issues

Risk issue	Risk factor (Hazard)	Risk domain class	Impact	Risk domain class
1	Hallucination Biased output	3-1, 3-2		
2	Unpredictable behavior	3-1	Serious damage due to misinformation in medical, legal, etc.	3-1, 3-2
3	Toxic contents creation Biased output	3-2	Social Stereotypes and Unfair Discrimination Hate speech and offensive terms Spreading false or misleading information	1-1, 1-2, 3-1, 3-2
4	Biased output	3-2	Social Stereotypes and Unfair Discrimination Reinforcement of social bias Fixation of misinformation and false information	1-2, 6-1, 6-2
5	Intentional Misinformation Creation Generating disinformation and propaganda	3-1 4-1	Social Stereotypes and Unfair Discrimination Exclusionary norm Spreading false or misleading information	1-2, 3-1, 3-2
6	Overly believe in generative AI	5	Fostering inappropriate use (reduced awareness of risks)	5
7	Information leakage Information estimation	2-1	Privacy infringement Security breach	2-1
8	Generating infringing data	2-2	Copyright infringement	2-2
9	Advancement of Automation by AI	6-1	Economic impact (e.g., replacement of workers)	6-3
10	Generating infringing data Support for attack code generation	2-2 4-2	Privacy infringement Security breach Facilitating fraud and targeted manipulation	2-1 4-2
11	Advancement of Automation by AI	6-1	Economic impact (e.g., replacement of workers)	6-3
12	Acceleration of technology development competition	6-1	Lowered safety standards and proliferation of bad norms	6-1
13	Increased power consumption during training and inference	6-3	Impact on natural environment	6-1
14	Spread of AI-produced information	6-2	Fixation of misinformation and false information	6-2
15	Generating disinformation and propaganda Overly believe in generative AI	4-1, 5	Reduced hurdles to malicious users Encouraging inappropriate use	4-1, 4-2, 5
16			Used for weapons proliferation	4-1
17			Illegal surveillance and censorship	4-1
18	Increase in size of training data	6-3	Lack of traceability Missing information on origin of training data	2-2 6-2
19	Interactions with other systems	4-2	Reduced hurdles to malicious users	4-1, 4-2
20	Lack of creativity in AI products	2-2	Failure of rights acquisition	2-2