Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
©2025 ESREL SRA-E 2025 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P9196-cd

Safe and Unsafe Information: Managing Risks in the Era of Generative Artificial Intelligence

Paolo Spagnoletti

Department of Information Systems, University of Agder, Norway. E-mail: paolo.spagnoletti@uia.no

Richard Baskerville

Computer Information Systems, Georgia State University, USA. E-mail: baskerville@acm.org

The transformative impact of digitalization on organizations has significantly increased the availability of organizational information to the public. This shift amplifies the responsibility of organizations to ensure the safety of their digital products and services, as unsafe information can cause harm to society or the environment. Generative artificial intelligence (GenAI) introduces unique risks by enabling the effortless production of ungrounded and potentially harmful content, such as hallucinations, which can propagate misinformation when uncritically used. These challenges necessitate a departure from traditional corporate social responsibility (CSR) frameworks towards more robust risk management strategies. This paper develops a taxonomy of characteristics of safe versus unsafe information from GenAI, characterized by three dimensions: correct, open, and benignant for safe information; and incorrect, protected, and dangerous for unsafe information. Drawing on empirical data from Italian organizations we validate and verify the alignment of established risk taxonomies and derive practical recommendations for mitigating these risks. These include implementing rigorous data validation pipelines, restricting inputs to trusted and verified sources, and employing robust processing and oversight mechanisms. By embedding these strategies into governance frameworks, organizations can mitigate the risks of unsafe information while ensuring that GenAI contributes positively to societal and environmental well-being.

Keywords: Data governance, Cybersecurity, Large Language Models, Botshit, Fake News.

1. Introduction

The rapid adoption of generative artificial intelligence (GenAI) systems has transformed the production and dissemination of organizational information. While these advancements hold great potential, they also introduce significant risks, particularly regarding the safety of the information produced (Rauh et al. 2024). Public use of this information, especially digital products and services, implies a large degree of organizational responsibility for their safety-in-use. Such safety means that this organizationally produced information should not cause harm to society or the environment. While CSR extends to help guide organizations in governing safety in the use of their public information, the capabilities of generative artificial intelligence enable easy production of botshit (Hannigan, McCarthy, and Spicer 2024; Hicks, Humphries, and Slater 2024): chatbot generated content that is not grounded in truth (i.e., hallucinations) and is then uncritically used in daily tasks. Such fake versions of information are beyond the control of the organization, yet potentially ungoverned and unsafe. Moreover, GenAI models are used maliciously to enhance, automate, and improve the execution of online fraud or cyberattacks (Gupta et al. 2023) or to create manipulated images, voices, texts, or videos, as well as fake profiles that interact with social media posts. While products like ChatGPT and Llama attempt to mitigate these risks by refusing to respond to certain prompts, the probabilistic nature of large language models (LLMs) means that jailbreaking techniques can bypass these safeguards, enabling access to harmful or dangerous outputs (Kim et al. 2024). Therefore, there is a new risk that otherwise safe digital products and services will be launched into a sea of botshit. CSR will not permit organizations to ignore such dangers. In this situation, organizations need new forms of risk management that provide provenance for their digital informational products and services. For example, the design and implementation of advanced data governance models would become a potential risk management tool for embedding safety in organizational information generated for public use.

In the realm of generative AI, risks associated with inaccuracies and the veracity of the outputs are particularly concerning. Although these models can generate content with notable accuracy, they are often referred to as "stochastic parrots" (Bender et al. 2021) because they lack a genuine understanding of meaning. The phenomenon of hallucinations (Ji et al. 2023) occurs when user requests fall outside the model's training data, leading to responses that, while seemingly plausible, are not grounded in reliable data. This can result in content that appears correct but is fundamentally erroneous (Hannigan et al., 2024), posing significant risks, particularly in safety- sensitive environments.

This paper explores the distinction between safe and unsafe information generated by GenAI, emphasizing safety as a critical quality of such information. By drawing parallels with well-established security and safety frameworks, the study delves into the challenges posed by GenAI in various applications and industries. Through an abductive analysis of empirical data collected from Italian companies and validated against existing taxonomies of risks, this research identifies the core characteristics of unsafe information: incorrect, protected from scrutiny, and dangerous. The findings contribute to a taxonomy of the characteristics of safe information: for GenAI this taxonomy regards its accuracy, openness, and benignant nature. Furthermore, the study presents actionable recommendations for organizations to mitigate the risks of unsafe information and align GenAI with CSR practices. This

work offers a foundation for designing advanced risk management strategies, bridging theoretical insights with practical applications, and ensuring that organizations can harness the benefits of GenAI while safeguarding their stakeholders and society at large.

2. Safe and Unsafe Information

The scope of this paper is characterizing safe versus unsafe information specifically, that which is produced from GenAI. While tightly related to GenAI systems, we regard safety as a quality of the information produced by GenAI systems. See Figure 1. A parallel metaphor would be safety of a motor vehicle trip. A car can produce travel. The car itself may be unsafe: bald tires or faulty brakes. Its use for travel is likely to be unsafe: a breakdown or a collision. But interest is in harms as a quality of the travel, not just the car.

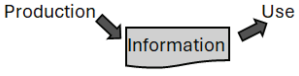

Figure 1. Information production and use.

Safety can often be confused or conflated with security. The two concepts can indeed be in some contexts (Piètre-Cambacédès and Chaudet 2010). For perhaps the most parsimonious distinction we adapt the inverse definitions found in Line et al. (2006). The quality of information security is the inability of the natural, technical, and social environment to affect the information in an undesirable way while information safety is the inability of the information to affect its natural, technical, and social environment in an undesirable way. Security is usually associated with intentionality, "the degree to which malicious harm is prevented, detected, and reacted to". Safety is usually associated with accidents, "the degree to which accidental harm is prevented, detected, and reacted to" (Firesmith 2003, p. viii).

Information security is often characterized as the protection of information confidentiality, integrity, and availability (the "CIA Triad"). While GenAI has brought many concerns for information safety to the fore, such security characteristics only partially apply. Certainly, the protection of information integrity is a concern for both security and safety. But, at best, security protection of confidentiality and availability apply differently for safety concerns.

2. Taxonomy Development (Empirical to Conceptual)

Nickerson, Varshney, and Muntermann (2013) describe a taxonomy development method based on a survey of the information systems literature. Our adaptation of their method is depicted in Figure 2.

Our concern is the quality of safety in the information produced by such systems. Our meta-characteristics are those that distinguish safe from unsafe information (Figure 2, step 1). These characteristics may differ from those of the information system that produced the information; and from those people or systems that may use the information. In the growing literature regarding the risks of GenAI, much of the literature deals with cases of risks of the use of AI in specific applications. Dominant cases include finance, cybersecurity, medicine, education, and management. There are also papers that consider risks to certain industries and countries, e.g., healthcare, automobile manufacturing, China, and Malaysia. Other cases include the dangers of specific GenAI applications in terms of deepfakes, home automation, and IoT.

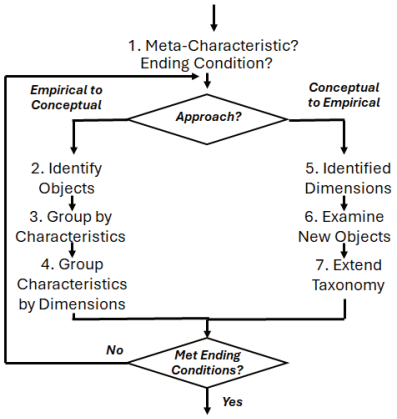

Figure 2.Adapted Taxonomy Development Method

Given the recency of the phenomenon, taxonomy studies of GenAI risks in general are rarer. There are four with good grounds for validity. NIST AI 600-1 (2024) provides a risk taxonomy developed by a group of experts with broad input from an interested community of experts. This document is a GenAI supplement (a "profile") to the more general NIST AI risk management framework (NIST AI 100-1 2024). The second study we identified is by Atkinson and Morrison (2024). They classify legal risks to GenAI developers based on prevalent claims in 29 lawsuits. In a third taxonomy, Beltran, Ruiz Mondragon, and Han (2024) base their work on government GenAI usage guidelines from five countries: Australia, Canada, New Zealand, the United Kingdom, and South Korea. Wach et al. (2023) provide a fourth taxonomy based on their analysis of literature review of both scholarly and practical articles from 2019-2023.

In addition to these four, we found two other useful taxonomies that were less well empirically grounded. Bengio et al. (2024) develop a risk taxonomy of general AI through an international summit of 30 national representatives; but this does not regard GenAI specifically. Tredinnick and Laybats (2023) provides an editorial opinion with another taxonomy of the dangers of GenAI; but this is based on opinion rather than an expert panel or analysis of sources.

These taxonomies are consistent but vary in criteria and level of analysis. Moreover, these are taxonomies of risks, rather than characteristics of safety. We abductively analyzed the four grounded taxonomies (with confirmation from the two additional ones) translating the risks into the characteristics of information safety that would be determined by such risks (these risks are our objects, Figure 2, step 2). We purposely sought utility: simple but reasonable dichotomous characteristics that would be both distinctive in definition and

exclusive of its opposite characteristic (our stopping condition, Figure 2, step 1). While there will always be grey zones, we depend on reasonable and useful distinctions.

We examined each of the risks in the four published taxonomies in identifying characteristics of the risks to GenAI information implied by specified risks of GenAI systems (Figure 2, step 3). Risks characterizing only systems (e.g., unauthorized access or environmental impact of compute resources) and risks characterizing only use (e.g., job displacement or technostress) were excluded. In subsequent iterations, we brought the remaining two studies into the analysis to check for any missing risks (none). The characteristics (originally five) were refined into three dimensions; mainly to eliminate issues with mutual exclusivity (Figure 4, step 4).

The result is a simple taxonomy (with plain English labels) of three defining characteristics of safe information from GenAI: the information is correct, open, and benignant (a logical conjunction). In contrast, there are three defining characteristics of unsafe information from GenAI: the information is incorrect, protected from disclosure, and/or dangerous (a logical sum).

Safe information is *correct*: "In accordance with fact, truth, or reason; free from error; exact, true, accurate; right." (OED 2024b) Unsafe information is *incorrect*: "Not in accordance with fact; erroneous, inaccurate" (OED 2024c). Grey zones are common: information that is partly correct and partly incorrect. In such cases the whole of the information must necessarily be regarded as unsafe: the body of it must be characterized as incorrect.

Safe information is *open*: "Not restricted to a few, generally accessible or available; such that anyone may use it, share it, or take part in it" (OED 2024d). Unsafe information is *protected*: sheltered, defended, or preserved from harm, danger, damage, etc. (OED 2024e). When GenAI is trained with such protected information, the GenAI may include some of this protected information in its generated outputs. Hence, such GenAI-produced information becomes unsafe because it contains protected information like trade secrets, copyrighted material, personally private information, etc. In the case of protected information, the legal structures governing use are usually based on negligence and tort.

Safe information is *benignant*: "Exerting a good or kindly influence; favourable, beneficial, salutary" (OED 2024a). Such information has many socially beneficial applications and fewer (or less extensive) socially harmful ones. Unsafe information is *dangerous*: "Fraught with danger or risk; causing or occasioning danger; perilous, hazardous, risky …" (OED 2023). Such information has many socially harmful applications and fewer (or less extensive) socially beneficial ones. When GenAI is trained with such dangerous information, the GenAI may include some of this dangerous information in its generated outputs. Hence, such GenAI-produced information becomes unsafe because it contains dangerous information like bomb-making instructions, state secrets, military secrets, contraband (such as child pornography), etc. In the case of dangerous information, the legal structures governing possession are based on criminal acts. There is a grey zone here because dangerous information might be regarded as an excessive form of protected information. The differences lie in how harms arise. With protected information, harms can arise in the use of such information, typically remedied by civil law; with dangerous information, harms can arise in the mere possession of such information (as well as its use), typically remedied by criminal law.

Table 1 presents examples of unsafe information content because these examples are taken from taxonomies of risks. By definitions, safe information is defined as the absence of such unsafe content. Consequently, safe information must have none of the unsafe content. Logically this is a conjunction (a logical "and"). Any unsafe information content will render the information unsafe. Logically this content is a logical sum (a logical "or").

3. Taxonomy Development (Conceptual to Empirical)

For our adaptation of the Taxonomy Development Method, we continued the process to validate and elaborate the taxonomy of unsafe information from GenAI (Figure 2, step 5) with abductive iterations (Sætre and Van de Ven 2021), leveraging empirical data collected from two distinct yet complementary sources to test the conceptual taxonomy from steps 1-4 (Figure 2). These empirical sources provide a robust foundation for analysing real-world challenges and risks associated with the integration of GenAI solutions in business. The empirical dataset comprises qualitative data obtained from two groups of Italian companies selected to reflect a range of experiences with GenAI technologies. The first group includes approximately 30 companies actively participating in a working group within Confindustria Servizi Innovativi e Tecnologici (CSIT), a leading organization representing over 5,500 companies and 200,000 employees in Italy's innovative and digital services sector. CSIT engages various industries, including financial services and IT. Established in 2024, the working group explores the risks and opportunities that GenAI presents for CSIT associates and their clients. Primary data were collected through semi-structured interviews with 15 key stakeholders, including 10 CEOs, 2 Chairs and 3 Chief Technology Officers (CTOs) from companies at the forefront of GenAI adoption. These companies, with an average age of 22 years and an average annual revenue of €5 million, represent a mix of established and mid-sized firms. The interviews, conducted between June and September 2024, totaled approximately 14 hours of discussion, with each session lasting between 50 and 70 minutes. The conversations focused on strategic insights into GenAI implementation, challenges, and future opportunities,. With participants' consent, all interviews were recorded for accurate transcription and analysis. Additionally, secondary data – including internal reports, industry analyses, and white papers shared by CSIT members - were incorporated to complement and validate the primary findings.

Table 1. Examples of unsafe information. The sources are studies of risk, so only examples of unsafe information are given. Safe information is free of such examples

Safe Info	Unsafe Info	Examples of unsafe information content	Sources of risk taxonomy examples
Correct	Incorrect	Errors, hallucinations, misrepresntations, scams, outdated material, biased material, or faked material (etc.)	(Beltran, Ruiz Mondragon, and Han 2024) (NIST AI 600-1 2024) (Wach et al. 2023) (Bengio et al. 2024) (Tredinnick and Laybats 2023)
Open	Protected	Intellectual property, copyrighted material, trade secrets, or private material (etc.)	(Atkinson and Morrison 2024) (Beltran, Ruiz Mondragon, and Han 2024) (NIST AI 600-1 2024) (Wach et al. 2023) (Bengio et al. 2024) (Tredinnick and Laybats 2023)
Benignant	Dangerous	Contraband or offensive material, state secrets, violent material, or weaponry material (etc.)	(Beltran, Ruiz Mondragon, and Han 2024) (NIST AI 600-1 2024) (Bengio et al. 2024) (Tredinnick and Laybats 2023)

The second group comprises companies identified by 60 students enrolled in the eighth edition of the "Master in Cybersecurity" organized by a leading University in Italy. This multidisciplinary Master's program tasked students with identifying Small and Medium Enterprises (SMEs) with experience in GenAI implementations and conducting risk management exercises. The resulting dataset includes 14 comprehensive project reports, each supervised by 3 instructors to ensure rigor and alignment with the program's academic standards. These reports provide practical, field-based perspectives on the risks of GenAI applications across various industries. A summary of projects is reported in Appendix A.

The abductive process to apply the taxonomy began with a comparative analysis of the collected data (Figure 2, step 6). Insights from both datasets were integrated to bridge high-level strategic viewpoints from the CSIT working group with granular, field-specific risk analyses from the Master students' reports. This combination added significant value by uniting institutional expertise with practical, hands-on investigations.

The analysis unfolded in three stages. First, risks associated with GenAI were identified through interviews, secondary data, and student reports. The initial focus was on risks recognized by industry leaders, which were validated through systematic field studies conducted by the Master students. Second, these identified risks were mapped against existing taxonomies of GenAI risks, such as those proposed by NIST AI 600-1 (2024), Atkinson and Morrison (2024), and Beltran et al. (2024). This mapping process translated the risks into defining characteristics of unsafe information. Finally, the findings were synthesized into a coherent taxonomy of unsafe information, emphasizing three core characteristics: incorrect, protected from disclosure, and dangerous. Validation was achieved by cross-referencing with grounded and supplementary taxonomies in the literature.

4. Elaborated Taxonomy

The analysis of qualitative data collected from interviews with CSIT-associated companies and the insights derived from their experiences with GenAI implementations reveal interesting insights and more detailed characteristics underlying the dimensions in the taxonomy of unsafe and safe information produced by GenAI. This section examines these findings through the lens of the taxonomy of safe information: correct, open, and benignant. Unsafe information, by contrast, is characterized as incorrect, protected from disclosure, and/or dangerous. Further characteristics, in the form of risk mitigation measures, implemented by organizations are also highlighted.

Table 2 summarizes the findings, aligning instances of unsafe information with their taxonomy characteristics, exemplary quotes, and risk mitigation measures employed (Figure 2, step 7).

From the analysis, it is evident that organizations adopting GenAI face significant risks when the information produced is incorrect, protected from critical scrutiny, or dangerous in its potential applications. The following risk mitigation strategies for unsafe information from GenAI emerge from our data:

- Data Validation Pipelines: organizations should invest in robust data preparation and validation processes. This includes adding metadata to documents, leveraging similarity criteria to validate generated content against trusted sources, and employing algorithms to ensure

outputs fall within acceptable ranges of parameters. These measures mitigate the risk of producing incorrect information by ensuring the reliability and accuracy of the underlying data and outputs (Case 13).

- Trusted and Verified Sources: organizations must limit the data inputs of GenAI systems to verified, authoritative sources. Certified databases, such as those validated by government or regulatory bodies, should be prioritized to prevent the generation of protected or unverified information (Case 9).
- Processing and Oversight Mechanisms: organizations should implement post-processing strategies and human oversight to avoid the production of dangerous information. Domain-specific rules can help validate outputs and block irrelevant or harmful queries. Human reviewers should also evaluate outputs in high-risk applications to ensure compliance with ethical and organizational standards (Cases 13 and 14).

Table 2. Examples of risk mitigation strategies for unsafe information from GenAI.

Unsafe Info	Related Quotes	Risk Mitigation Strategies
Incorrect	Case 13: "If I use a generative model for augmented research and ask, for example, at what pressure car tires should be inflated, the system cannot respond '1,500 bars' due to an error or hallucination. It must provide a value consistent with reality, because inflating tires to 1,500 bars would make them explode and could cause accidents."	Implementing checks to ensure outputs fall within acceptable ranges of parameters and rejecting outputs that deviate significantly from established norms. Algorithms are employed to qualify results and discard implausible data.
	Case 6: "Of course, hallucinations exist, but what we have done is not so much eliminate hallucinations—which is impossible—but ensure that the product alerts you when there is a risk of them occurring."	Developing systems to flag potential hallucinations by distinguishing between responses based on internal model knowledge versus external, validated documents.
Protected	Case 9, Quote 1: "Let me give you a simple example: the pharmaceutical database and all related information about medicines. These are reliable sources, validated by the Ministry of Health or pharmaceutical companies. In these cases, ChatGPT or any other AI cannot improvise. If the AI works only on these verified sources, then yes, it can produce useful and reliable content. This is the crucial point: AIs must work only on certain sources."	Restricting GenAI's input to verified and authoritative data sources, such as certified databases, to ensure accuracy and reliability of generated information.
Dangerous	Case 13: "If, for example, I ask for information about the maintenance operations for my car, the system cannot respond with information about a train from Modena to Milan. We need to block these out-of-context questions, which a generative model might otherwise attempt to answer."	Implementing domain-specific rules to block out-of-context queries and using contextual understanding technologies to exclude irrelevant requests.
	Case 14: "You can set prompts to prepare the context or provide preliminary information. Finally, you can do post-processing on the responses. After the system provides an answer, you can apply reasoning systems or specific rules, using a neuro-symbolic approach."	Post-processing of outputs using neuro-symbolic reasoning systems to validate responses and ensure compliance with organizational policies and ethical guidelines.

5. Conclusion

Our study shows the applicability of the taxonomy of safe information to design risk management strategies in the deployment of GenAI. As implication for practice, organizations must go beyond traditional CSR frameworks to address the unique risks posed by GenAI. This requires the implementation of advanced risk management strategies, including data validation pipelines, the use of trusted and verified sources, and robust processing and oversight mechanisms. By embedding these strategies into their operational and governance frameworks, organizations can mitigate the risks associated with unsafe information while fostering the production of information that is correct, open, and benignant. In doing so, organizations can navigate the complex landscape of generative AI, ensuring that their digital products and services contribute positively to society and the environment. This study underscores the need for proactive and adaptive risk management practices, setting a foundation for future research and practical advancements in the safe deployment of GenAI.

References

Atkinson, David, and Jacob Morrison. 2024. "A Legal Risk Taxonomy for Generative Artificial Intelligence." *arXiv preprint arXiv:2404.09479*.

Beltran, Marco Antonio, Marina Ivette Ruiz Mondragon, and Seung Hun Han. 2024. "Comparative analysis of generative ai risks in

the public sector." Proceedings of the 25th Annual International Conference on Digital Government Research.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada. https://doi.org/10.1145/3442188.3445922.

Bengio, Y., S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, D. Goldfarb, H. Heidari, L. Khalatbari, S. Longpre, V. Mavroudis, M. Mazeika, K. Y. Ng, C. T. Okolo, D. Raji, T. Skeadas, F. Tramèr, B. Fox, A. C. P. de Leon Ferreira de Carvalho, M. Nemer, R. Pezoa Rivera, Y. Zeng, J. Heikkilä, G. Avrin, A. Krüger, B. Ravindran, H. Riza, C. Seoighe, Z. Katzir, A. Monti, H. Kitano, M. Kerema, J. R. López Portillo, H. Sheikh, G. Jolly, O. Ajala, D. Ligot, K. M. Lee, A. H. Hatip, C. Rugege, F. Albalawi, D. Wong, N. Oliver, C. Busch, O. Molchanovskyi, M. Alserkal, S. M. Khan, A. McLean, A. Gill, B. Adekanmbi, P. Christiano, D. Dalrymple, T. G. Dietterich, E. Felten, P. Fung, P.-O. Gourinchas, N. Jennings, A. Krause, P. Liang, T. Ludermir, V. Marda, H. Margetts, J. A. McDermid, A. Narayanan, A. Nelson, A. Oh, G. Ramchurn, S. Russell, M. Schaake, D. Song, A. Soto, L. Tiedrich, G. Varoquaux, A. Yao, and Y.-Q. Zhang. 2024. International Scientific Report on the Safety of Advanced AI: Interim Report. edited by AI Seoul Summit: UK Department for Science, Innovation & Technology.

Firesmith, Donald G. 2003. *Common concepts underlying safety, security, and survivability engineering.* Carnegie Mellon University, Software Engineering Institute Pittsburgh, Pa, USA.

Gupta, M., C. Akiri, K. Aryal, E. Parker, and L. Praharaj. 2023. "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy." *IEEE Access* 11: 80218-80245. https://doi.org/10.1109/ACCESS.2023.3300381.

Hannigan, Timothy R., Ian P. McCarthy, and André Spicer. 2024. "Beware of botshit: How to manage the epistemic risks of generative chatbots." *Business Horizons* 67 (5): 471-486. https://doi.org/https://doi.org/10.1016/j.bushor.2024.03.001.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. "ChatGPT is bullshit." *Ethics and Information Technology* 26 (2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55 (12): Article 248. https://doi.org/10.1145/3571730.

Kim, Minseon, Hyomin Lee, Boqing Gong, Huishuai Zhang, and Sung Ju Hwang. 2024. "Automatic Jailbreaking of the Text-to-Image Generative AI Systems." *arXiv preprint arXiv:2405.16567*. https://doi.org/https://doi.org/10.48550/arXiv.2405.16567.

Line, Maria B, Odd Nordland, Lillian Røstad, and Inger Anne Tøndel. 2006. "Safety vs security?" Proceedings of the 8th International Conference on Probabilistic Safety Assessment and Management, PSAM 2006, New Orleans, USA.

Nickerson, Robert C., Upkar Varshney, and Jan Muntermann. 2013. "A method for taxonomy development and its application in information systems." *European Journal of Information Systems* 22 (3): 336-359. https://doi.org/10.1057/ejis.2012.26.

NIST AI 100-1. 2024. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST Gaithersburg, MD, USA.

NIST AI 600-1. 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST Trustworthy and Responsible AI Gaithersburg, MD, USA.

OED. 2023. dangerous, adj., sense 2. In *Oxford English Dictionary*: Oxford University Press.

---. 2024a. benignant, adj., sense 2.a. In *Oxford English Dictionary*: Oxford University Press.

---. 2024b. correct, adj., sense II.3. In *Oxford English Dictionary*: Oxford University Press.

---. 2024c. incorrect, adj., sense 4. In *Oxford English Dictionary*: Oxford University Press.

---. 2024d. open, adj., sense II.28.a. In *Oxford English Dictionary*: Oxford University Press.

---. 2024e. protection, n., sense 1.a. In *Oxford English Dictionary*: Oxford University Press.

Piètre-Cambacédès, Ludovic, and Claude Chaudet. 2010. "The SEMA referential framework: Avoiding ambiguities in the terms "security" and "safety"." *International Journal of Critical Infrastructure Protection* 3 (2): 55-66. https://doi.org/https://doi.org/10.1016/j.ijcip.2010.06.003.

Rauh, Maribeth, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. 2024. "Gaps in the Safety Evaluation of Generative AI." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (1): 1200-1217. https://doi.org/10.1609/aies.v7i1.31717.

Sætre, Alf Steiner, and Andrew Van de Ven. 2021. "Generating Theory by Abduction." *Academy of Management Review* 46 (4): 684-701. https://doi.org/10.5465/amr.2019.0233.

Tredinnick, Luke, and Claire Laybats. 2023. "The dangers of generative artificial intelligence." *Business Information Review* 40 (2): 46-48. https://doi.org/10.1177/02663821231183756.

Wach, Krzysztof, Cong Doanh Duong, Joanna Ejdys, RūtaRuta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. 2023. "The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT." *Entrepreneurial Business and Economics Review* 11 (2): 7-30. https://doi.org/https://doi.org/10.15678/EBER.2023.110201.

Appendix A

ID	Project	Description	Unsafe Information Characteristic
1	Generative AI for Data Synthesis: Innovation Opportunities and Security Challenges	Analyzes risks and opportunities in GAI adoption for processing and synthesizing sensitive documents.	Wrong, Protected
2	'AI m fighter': Generative AI in the Fitness Sector	Explores personalized training programs for athletes, addressing cyber risks.	Dangerous
3	The Faba Project - Storytelling for Children with Generative AI: Risks and Safeguards	Focuses on creating narratives for children, analyzing data privacy and ethical challenges.	Protected, Dangerous
4	Generative AI and Cybersecurity: An Integration Model for Apulia Soft	Studies automation benefits and risks in a software development context.	Wrong, Dangerous
5	Revolutionizing Learning with Generative AI: The Smart Notes Model	Examines a tool for automating note creation during training events.	Wrong
6	Introduction of Generative AI Tools in a Laboratory of Analysis and Diagnostic Imaging: Opportunities and Risk and Compliance Management	Analyzes AI in medical diagnostics, focusing on privacy and compliance.	Wrong, Dangerous
7	Opportunities and Risks in the Implementation of Generative AI in SMEs: A Case Study on AI-Travel Assistant	Examines ethical and legal challenges in implementing AI in a travel agency.	Protected, Dangerous
8	Wonderflow: The Digital Bridge Between Companies and Consumers	Analyzes AI to synthesize consumer reviews while addressing privacy issues.	Protected
9	Secure Code Development with the Aid of Generative AI: Risks and New Opportunities	Focuses on AI for secure code development, addressing privacy and compliance.	Wrong, Protected
10	Generative Artificial Intelligence at the Service of Energy: The case of Electrade - Bitdrop	Studies AI in energy optimization, addressing data security and privacy.	Protected, Dangerous
11	Use of Generative AI for Training Predictive Models: Safety and Assistance in Diabetic Pathology	Investigates GANs for training predictive models in diabetes management.	Wrong, Dangerous
12	Generative Artificial Intelligence in Human Resources Management: Risks and Guidelines for CV Selection	Analyzes risks of bias and data privacy in automating CV analysis.	Protected, Dangerous
13	The Double Face of Generative AI: the Cy4Gate case study between business innovation and risks	Examines secure integration of AI in regulated business environments.	Wrong
14	The implementation of generative AI in diagnostic medical applications: Gleamer BoneView and the analysis of privacy and security risks of synthetic data	Analyzes GANs for privacy and security in medical imaging.	Wrong, Protected