AI Security Map: Holistic Organization of AI Security Technologies and Impacts on Stakeholders

Hiroya Kato
KDDI Research, Inc.
Fujimino, Saitama, Japan

Kentaro Kita
KDDI Research, Inc.
Fujimino, Saitama, Japan

Kento Hasegawa
KDDI Research, Inc.
Fujimino, Saitama, Japan

Seira Hidano
KDDI Research, Inc.
Fujimino, Saitama, Japan

Abstract

As the social implementation of AI has been steadily progressing, research and development related to AI security has also been increasing. However, existing studies have been limited to organizing related techniques, attacks, defenses, and risks in terms of specific domains or AI elements. Thus, it extremely difficult to understand the relationships among them and how negative impacts on stakeholders are brought about. In this paper, we argue that the knowledge, technologies, and social impacts related to AI security should be holistically organized to help understand relationships among them. To this end, we first develop an AI security map that holistically organizes interrelationships among elements related to AI security as well as negative impacts on information systems and stakeholders. This map consists of the two aspects, namely the information system aspect (ISA) and the external influence aspect (EIA). The elements that AI should fulfill within information systems are classified under the ISA. The EIA includes elements that affect stakeholders as a result of AI being attacked or misused. For each element, corresponding negative impacts are identified. By referring to the AI security map, one can understand the potential negative impacts, along with their causes and countermeasures. Additionally, our map helps clarify how the negative impacts on AI-based systems relate to those on stakeholders. We show some findings newly obtained by referring to our map. We also provide several recommendations and open problems to guide future AI security communities.

CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

AI Security, Privacy, Compromise of AI, Misuse of AI, Social impact

1 Introduction

As AI is increasingly utilized in society, research and development on AI security is also further accelerating. AI security is not only closely connected to the traditional elements of information security, namely confidentiality, integrity, and availability (CIA) but is also strongly related to other AI elements, such as explainability and fairness. Accordingly, concerns are growing regarding the complex negative impacts on stakeholders such as individuals and society. In other words, the scope of AI security has now extended beyond AI itself to encompass individuals and society, and these are also closely interconnected. Nevertheless, most existing survey papers [8, 15, 27, 36, 60] and systematization-of-knowledge (SoK) papers [9, 23, 29, 56] have been limited to the classification of attack and defense techniques for AI. Some studies systematically investigate techniques or risks in terms of specific AI elements such as explainability [11, 26, 41], fairness [19, 35, 61], and privacy [7, 12, 22, 37], and examine real-world risks or social impacts [2, 6, 33, 40, 46, 55]. However, the relationships among multiple elements including those on AI security have not comprehensively been organized, and they are exclusively within limited domains such as LLMs. In other words, most existing studies have been limited to organizing related techniques, attacks, defenses, and social impacts from distinct perspectives of fields or AI elements. Our position is that **the knowledge, technologies, and social impacts related to AI security should be holistically organized to help understand relationships among them**. To the best of our knowledge, there is no study that holistically organize the relationships between AI elements or provide an inclusive classification that covers negative impacts on stakeholders.

Thus, in this paper, we develop an AI security map that holistically organizes these relationships through examining these interconnections and impacts. This map consists of the two aspects, namely the information system aspect (ISA) and the external influence aspect (EIA). The elements that AI should fulfill within information systems are classified under the ISA. On the other hand, the EIA includes elements that affect individuals and society. Our map organizes the elements related to AI security, as well as how individuals and society can be affected as a result of AI being attacked or misused. To develop our map, we surveyed existing papers to classify and organize attacks that can compromise the CIA. Furthermore, we organized other elements affected by the compromise of the CIA and considered negative impacts that may arise from the ISA. Similarly, we considered elements in the EIA on the basis of the negative impacts in the ISA and existing papers. For each element in the EIA, we organized negative impacts caused by the compromise or misuse of AI. Finally, attacks, causal factors, defenses, and related elements are mapped to each negative impact. By referring to our map, one can understand the potential negative impacts on AI-based information systems, along with defense and countermeasures necessary to prevent such impacts. Additionally, our AI security map helps clarify how these negative impacts on the information systems influence individuals and society. Our contributions are as follows:

- We holistically organize not only multiple elements related to AI, but also those impacting individuals and society.
- We consider the mutual relationships between the ISA and the EIA, clarifying how negative impacts on individuals and society are caused by those on information systems.
- Our map organizes not only negative impacts caused by attacks on AI but also ones resulting from AI misuse.
- In contrast to prior work that has primarily focused on real-world risks or social impacts within specific domains, our work considers them in the broader context of AI security as a whole. In particular, we also focus on the subsequent influence resulting from particular risks or impacts.

2 Current Landscape of AI Security

As AI technologies are incorporated in various fields, concepts of AI security are diversified and become complex more and more. In such situations, researchers are working on various research directions, such as privacy and explainability, related to AI security. To present a systematic critical review on huge studies, main approaches, and evaluation methods in these areas, there are many survey papers and SoK papers. In what follows, we briefly introduce existing survey and SoK papers on these related research fields.

2.1 Attack and Defense

In the domain of AI security, there are a huge number of studies regarding attacks and defenses. Since providing taxonomies or systematic reviews of attacks and defenses is a primary way to realize clear overviews, existing survey papers [8, 15, 27, 36, 60] and SoK papers [9, 23, 29, 56] mainly focus on organizing attacks on AI or defenses. We consider that there are three types of taxonomies in existing studies classifying attacks and defenses.

Taxonomy in AI systems. The first type classifies attacks and defenses in AI systems. AI systems mean information systems using AI. For example, to represent an overview for AI security, Hu et al. [15] review the challenges and research advances for security issues in AI. In that work, the lifecycle of an AI system is used as a guide so as to introduce the security threats that emerge at each stage and corresponding countermeasures. Also, Chen and Babar [8] present a review regarding the security for machine learning (ML) based software systems in light of the fact that no literature review is aimed at a comprehensive investigation of ML-based software systems from the secure development aspect. Kiribuchi et al. [21] provide an overview of adversarial attacks on AI systems. They identify 11 major attack types and their resulting impacts. One of the distinguishing features of that work is focusing on the impacts of attacks on AI systems and providing intuitive visual summaries.

Taxonomy in specific attacks. The second type provides a taxonomy with focus on specific attacks. There are studies that present a comprehensive summary in a specific type of attack on AI, such as poisoning attacks [36], backdoor attacks [60], and jailbreak attacks [24]. Ramirez et al. [36] conduct a survey with highlighting the most relevant information related to poisoning attacks. That work compiles the most relevant insights and findings found in the latest existing literature regarding poisoning attacks and several defense techniques. Zhang et al. [60] present a comprehensive and systematic summary of both backdoor attacks and defenses targeting multi-domain AI models so as to systematically analyze shortcomings of existing research and address the lack of comprehensive reviews. Lu et al. [24] propose AutoJailbreak, a framework designed to comprehensively evaluate the resilience of LLMs against jailbreak attacks. Furthermore, that work conducts a comprehensive examination of jailbreak attacks and defenses. In total, over 28 jailbreak attacks and 12 jailbreak defenses are organized.

Taxonomy in specific domains. The third type deals with the classification of attacks and defenses in specific domains. There are studies [27, 43, 56] that summarize related work in order to provide taxonomy in specific domains. Shen et al. [43] perform the first systematization of knowledge of growing AI security research in the field of autonomous driving (AD). They collect and analyze 53 existing papers, and systematically taxonomize them on the basis of research that is critical for the security field. That work organizes knowledge from the perspective of important research aspects, such as AI components related to the attack and defense as well as evaluation methodologies. To address the existing gaps in understanding the complexities of Edge AI, Wingarz et al. [56] provide a comprehensive survey of the challenges necessary for enhancing the security and safety of Edge AI. They examine both existing threats and their relevant countermeasures with social implications. Finally, they identify a series of open research challenges and present required action to advance solutions in the area.

Observation

Most existing studies just provide taxonomies of attacks and defenses from the perspective of individual research. Thus, there has been no description of what specific negative impacts in the real world can be caused by each attack.

2.2 Elements related to AI Security

In recent years, as AI is expected to be applied in various fields, AI related elements, such as fairness and privacy, have become increasingly important. It is crucial to consider whether AI meets these elements closely related to AI security because they can be compromised by attacks on AI, which may have significant negative impacts. Existing studies have summarized the various risks, challenges, and desired research directions in terms of each element entailed by the utilization of AI. In what follows, we introduce related studies that address representative elements in AI.

Explainability. As AI increasingly exerts an influence on human decision-making, it is becoming essential to understand the rationale behind AI predictions. To this end, Explainable AI (XAI) has been studied. XAI refers to technologies that enable the explanations of the reasoning and processes by which AI models make decisions or predictions in a manner comprehensible to humans. The advancement of XAI is expected to enhance the trustworthiness and transparency of AI predictions. There are various systematic studies [11, 26, 41] on the field of XAI. For example, Dwivedi et al. [11] survey programming techniques for XAI and present the different phases of XAI in a typical ML development process. The various XAI approaches are classified to discuss the key differences among the existing XAI techniques. Schwalbe and Finzel [41] provide a complete taxonomy of XAI methods with respect to notions included in the existing research.

Fairness. To prevent AI models from perpetuating the biases present in the data and producing unfair decisions, researchers has worked on the research on fairness in AI. Although fairness is a relatively new research area, as of now, there are multiple studies [19, 35, 61] in this area. For example, Zhang [61] review recent advances in AI fairness aimed at bridging gaps for practical deployment in real-world scenarios. That review seeks not only to identify existing gaps but also to propose solutions that reconcile the theoretical underpinnings of fairness with the complex realities of real-world data dynamics. Moreover, they highlight the limitations and significant potential for real applications. Parraga et al. [35] provide an in-depth overview of representative debiasing methods for fairness-aware neural networks in vision and language domains.

Privacy. Privacy is one of the most related domains to AI security because it is a key principle for developing ethical and secure AI. In studies [12, 22], AI privacy risks are presented. Golda et al. [12] conduct a meticulous examination of the privacy and security challenges inherent to Generative AI. That study provides five pivotal perspectives essential for a comprehensive understanding of these intricacies. Lee et al. [22] present 12 high-level privacy risks that AI technologies either newly created or exacerbated. There are also papers that systematically organize privacy risks or provide taxonomy of attacks. Chang et al. [7] provide a systematic overview of attacks on healthcare AI to facilitate privacy leakage and defenses in response to inconsistent settings in terms of healthcare deployment scenarios and threat models. Rigaki and Garcia [37] propose an privacy attack taxonomy, together with a threat model that allows the categorization of different attacks based on the adversarial knowledge. They also offer an overview of defenses and discussion of the open problems and future directions.

Observation

Existing studies systematically investigate techniques or risks in terms of specific AI elements. However, the relationships among multiple elements including those on AI security have not systematically been organized.

2.3 Real-World Risks and Impacts

As another direction, some studies also discuss what risks and impacts [33, 46, 55] are brought about by the compromise or misuse of AI [2, 6, 40] in the real world.

Social impact of AI. Weidinger et al. [55] develop a comprehensive taxonomy of ethical and social risks associated with language models (LMs). They identify 21 risks appearing in current LMs and develop a taxonomy consisting of 6 risk areas to help understand their landscape. That work shares foresight to help make the landscape of risks associated with LMs easier to parse, which contributes to guiding required action to address these risks. Slattery et al. [46] present several implications for the collective understanding of how the landscape of AI risks is constructed. That study highlights the need for a balanced approach that both drives technological progress and embraces social responsibility by considering the impact of AI on the workforce, economic dynamics, and ethical issues. Pankajakshan et al. [33] discuss LLM security and stakeholder risks. They argue that organizations are deploying LLM-integrated systems without understanding the severity of potential consequences. Moreover, whereas existing studies by OWASP and MITRE offer a general overview of threats and vulnerabilities, they pointed out that there is a few methods for directly and succinctly analyzing the risks for security practitioners, developers, and key decision-makers who are working with this novel technology. To address the limitations, they finally propose a risk assessment process.

Misuse of AI. AI can have adverse impacts on people and society regardless of attacks on AI. In addition to the attack on AI, the misuse of AI is one of the factors that can have negative impacts on the real world. This concern is discussed in several studies [2, 6, 40]. Aïmeur et al. [2] mention that fake news can have a significant impact on society, and false content is easier to generate and harder to detect by using AI based tools. Recent studies [6, 40] discuss a emerging paradigm that integrates AI technologies to conduct or enhance cyber attacks, which is called offensive AI (OAI). Castagnaro et al. [6] explore whether AI can enhance the directory enumeration process and propose a novel LM-based framework. On the other hand, Schröer et al. [40] devise a common set of criteria reflecting essential technological factors related to OAI. They consider OAI as the crucial means to violate security and privacy. In their studies, humans and society are considered and summarized as targets of OAI attacks.

Observation

Some studies discuss impacts on stakeholders beyond merely classifying attacks and defenses. However, they discuss real-world risks or social impacts exclusively within limited domains such as LLMs and OAI.

2.4 Limitation on Current Landscape

Most systematic studies on AI security have primarily organized knowledge from a technical perspective by classifying attacks on AI and defense methods. On the other hand, techniques or some risks related to AI elements are also discussed in existing studies. However, they focus on individual elements, such as fairness and privacy. Thus, the relationships between the elements and AI security have not been discussed or organized despite their close and important relationship. Some papers discuss impacts on stakeholders or society in addition to the classification of attacks and defenses. However, their overviews have been limited to real-world risks or impacts only in specific fields. Furthermore, the subsequent influence resulting from particular risks or impacts have not been considered. It is important to consider such chains of impacts so as to understand potential impacts on people and society in practice. Today, AI has a diverse impacts not only on information systems, but also on stakeholders. Consequently, the scope of AI security has expanded beyond a limited group of researchers to encompass most people including engineers, users, and even non-users. In a nutshell, it is no longer sufficient to examine individual elements in isolation, which requires to take the broader context into account.

3 Holistic Organization of AI Security

3.1 Position Statement

As described in the previous section, most existing studies have been limited to organizing related techniques, attacks, defenses, and social impacts from distinct perspectives or AI elements. Hence, it is extremely difficult to understand the relationships among related elements by referring to multiple papers in their current state. As AI becomes increasingly integrated into information systems and society, a holistic overview is required not only from the standpoints of research but also from those of business and general public in order to comprehensively understand potential impacts caused by the compromise or misuse of AI. In particular, in this increasingly complex field, interrelationships among all elements associated with AI security should be comprehensively organized. Furthermore, it is essential to clarify how negative impacts on individuals and society are brought about. In short, our position is summarized as follows:

> **Position**
>
> The knowledge, technologies, and social impacts related to AI security should be holistically organized to help understand relationships among them. To this end, it is necessary not only to classify attacks and defenses, but also to organize the interrelationships among related elements, as well as how individuals and society may be influenced by attacks or the misuse of AI.

In contrast to the previous studies, our work considers and summarizes the interrelationships between those elements through negative impacts, in addition to the classification of various types of attacks and defenses. In what follows, we propose new holistic organization that encompasses AI security technologies and negative impacts on information systems and stakeholders.

3.2 AI Security Map

We develop new holistic organization of AI security technology and impacts on information systems and stakeholders, which is called AI security map. This map consists of the two aspects, namely ISA and EIA. The elements that AI should fulfill within information systems are classified under the ISA. On the other hand, the EIA includes elements that affect individuals and society as a result of AI being attacked or misused. The definitions of the elements in the ISA and the EIA are shown in Table 6 and Table 7, respectively, in Appendix A. For each element, corresponding negative impacts are identified. The negative impacts in the ISA are mainly attributed to attacks on AI. They strongly related to information systems using AI. For example, confidentiality is breached when information systems that a LLM is integrated with are attacked via the prompt injection in order to elicit personal information from LLMs. On the other hand, the negative impacts categorized under the EIA may arise not only when AI is attacked but also when AI that is functioning properly is misused by malicious users. This map organizes these negative impacts, the attacks and factors that cause them, as well as the corresponding defense methods and countermeasures. In addition, we holistically examine the relationships between the ISA and the EIA, clarifying how the compromise or misuse of AI can bring about negative impacts on stakeholders such as individuals and society. We assume four types of security targets at this stage. These security targets mean primary stakeholders that could be affected by AI attacks or misuse. The definitions of these security targets are shown in Table 8 in Appendix B.

Information system aspect. In this aspect, the elements that AI must satisfy within information systems are classified. The primary focus is on the three elements of information security, known as CIA. Other elements are mainly organized based on their being affected by the compromise of CIA. By examining the negative impacts on these elements, one can better understand their underlying causes and potential countermeasures. Additionally, it is assumed that the negative impacts on the elements classified under this category may influence elements in the EIA.

External influence aspect. In this aspect, elements that impact individuals and society, such as privacy violations and infringements of rights like copyright, are classified. The negative impacts in this aspect are assumed to arise not only from attacks on AI within information systems but also from the misuse of AI. By examining the negative impacts on the elements within this aspect, one can understand which aspects of the information system are compromised and how such damage can affect people and society, as well as potential countermeasures.

Relationship of elements in AI security map. The AI security map organizes the relationships between the elements of the ISA and the EIA. Specifically, we consider that compromise or misuse of any element within the ISA is related to elements of the EIA. Elements within the ISA other than the CIA are affected by breaches of the CIA. Elements of the EIA are mainly impacted by compromise of elements within the ISA. By referring to the AI security map, one can holistically understand these relationships.

4 Discussion

In this section, we discuss new findings by referring to the AI security map. To be specific, we conduct discussion in order to answer the following questions.

(1) How are elements or negative impacts in the ISA interrelated?
(2) How are elements or negative impacts in the EIA interrelated?
(3) What impact is produced on each security target?
(4) How does negative impacts affect individuals and society?

At this stage, we represent the AI security map in the tabular form. Due to space limitations, the AI security map in terms of the ISA is divided into two parts: one for CIA-related elements and the other for non-CIA elements. Additionally, as for the EIA, we divide the map into three ones on the basis of security targets. In what follows, we show the concrete maps to provide some insights obtained by referring to them.

4.1 Insights into Information System Aspect

Table 1 shows the AI security map for negative impacts related to CIA in the ISA. We identified that the compromise of CIA elements causes 13 negative impacts on information systems. These negative impacts can directly hinder the functions and operations of AI based

Table 1: AI security map for negative impacts related to CIA in the ISA.

Elements	Negative impacts	Attack or cause	Defenses or countermeasures	Related elements in the EIA	
Confidentiality	Training data leakage	- Membership inference attack [45]	- Differential privacy (DP) [1] - Encryption technology [18] - AI access control	- Privacy - Copyright and authorship - Reputation	- Psychological impact - Compliance with laws and regulations
	Personal information leakage	- Membership inference attack [45] - Prompt injection [44]	- DP [1] - Federated learning [52] - Personal information masking - AI access control	- Privacy - Copyright and authorship - Reputation	- Psychological impact - Compliance with laws and regulations
	Reconstruction of training data	- Model inversion attack [62]	- DP [1] - Encryption technology [18] - AI access control	- Privacy - Copyright and authorship - Reputation	- Psychological impact - Compliance with laws and regulations
	Model information leakage	- Model extraction attack [50]	- DP [1] - Detection of model extraction attack - AI access control	- Privacy - Copyright and authorship - Reputation	- Psychological impact - Compliance with laws and regulations
	Leakage of system prompts	- Prompt leaking [17]	- Prompt checking	- Reputation	
Integrity	Manipulation of AI output for specific inputs	- Adversarial examples [34, 48]	- Adversarial training [25] - Detection of adversarial examples [49] - Certified robustness (CR) [23]	- Reputation - Disinformation - Usability - Consumer fairness	- Compliance with laws and regulations - Critical infrastructure - Physical impact - Medical care
	Degradation of AI performance due to training data contamination	- Poisoning attack [4, 5]	- Detection of poisoned data - CR [23]	- Reputation - Misinformation - Usability - Consumer fairness	- Compliance with laws and regulations - Critical infrastructure - Physical impact - Medical care
	Manipulation of AI output under specific conditions	- Backdoor attack [14, 38]	- Detection of triggers [10] - Detection of backdoored models [57] - CR [23]	- Privacy - Disinformation - Usability - Consumer fairness - Reputation - Human-centric principle	- Compliance with laws and regulations - Physical impact - Ethics - Economy - Critical infrastructure - Medical care
	Generation of harmful responses	- Prompt injection [44]	- Toxicity detection [16] - Prompt checking	- Privacy - Disinformation - Consumer fairness	- Reputation - Compliance with laws and regulations - Ethics
Availability	Misclassification by AI, leading to degradation of functionality or service quality	- Adversarial examples [34, 48]	- Adversarial training [25] - Detection of adversarial examples [49] - CR [23]	- Reputation - Human-centric principle - Ethics - Critical infrastructure	- Compliance with laws and regulations - Physical impact - Economy - Medical care
	Continuous decrease in predictive accuracy, leading to degradation or cessation of functionality or service quality	- Poisoning attack [4, 5]	- Detection of poisoned data - CR [23]	- Reputation - Usability - Physical impact - Psychological impact	- Financial impact - Economy - Critical infrastructure - Medical care
	AI output manipulated under specific conditions, leading to degradation of functionality or service quality	- Backdoor attack [14, 38]	- Detection of triggers [10] - Detection of backdoored models [57] - CR [23]	- Reputation - Usability - Physical impact - Psychological impact	- Financial impact - Economy - Critical infrastructure - Medical care
	Service disruption due to high consumption of AI resources	- Model DoS [59]	- Token limit - AI access control	- Reputation - Physical impact - Psychological impact - Financial impact	- Economy - Critical infrastructure - Medical care

information systems. Furthermore, we discovered that there are more negative impacts related to confidentiality compared to those on integrity and availability. This is consistent with the fact that research regarding privacy is actively conducted. In many cases, the elements of CIA are compromised first by attacks on AI. As a result, this can then affect other elements in the ISA.

Table 2 shows the AI security map for negative impacts related to elements other than CIA in the ISA. As to elements other than CIA, we have organized the six elements, namely explainability, output fairness, safety, accuracy, controllability, and trustworthiness. Among them, some elements can be caused by specific attacks. However, they can also occur as a result of the compromise of CIA elements. We identified one negative impact on each of these six elements. In particular, it is clear that the compromise of integrity can lead to negative impacts on all the six elements. For example, if integrity is attacked and the expected prediction results are no longer returned, this can affect explainability, making it impossible to provide explanations for the AI's prediction results. From the information security perspective, it is natural that integrity is related to various elements of information systems. As to trustworthiness, explainability can also bring about the negative impact. Overall, in the context of AI security, our map made us realize that integrity is a crucial element that affects many other aspects in information systems with AI.

Takeaway

Confidentiality tends to be the primary targets of attacks in AI based information systems. Also, integrity is the most influential elements in the ISA. Keeping integrity intact is extremely difficult but essential, preventing other elements from being compromised.

4.2 Insights into External Influence Aspect

In total, we identified 20 elements and 36 negative impacts in the EIA. This indicates that AI security is closely related to a variety of elements and can have a significant impact on individuals and

Table 2: AI security map for negative impacts related to elements other than CIA in the ISA.

Elements	Negative impacts	Compromised elements	Attack or cause	Defenses or countermeasures	Related elements in the EIA
Explainability	Difficulty in understanding AI inference results	- Integrity	- Attacks on explainability	- XAI [41] - Robust explainability	- Reputation - Transparency - Psychological impact - Financial impact - Economy - Medical care
Output Fairness	Bias in the AI output	- Integrity	- Bias in training data [42]	- Defensive method for integrity - Detection of bias in AI output - Elimination of bias in training data - Creation of fair AI models	- Usability - Consumer fairness - Reputation - Medical care - Compliance with laws and regulations - Psychological impact - Ethics
Safety	Harm to humans due to decreased prediction accuracy or unexpected behavior by AI	- Integrity		- Defensive method for integrity - Fail-safe mechanism	- Reputation - Human-centric principle - Physical impact - Psychological impact - Critical infrastructure - Compliance with laws and regulations - Financial impact - Ethics - Economy - Medical care
Accuracy	Decrease in AI prediction accuracy	- Integrity		- Defensive method for integrity	- Misinformation - Usability - Consumer fairness - Reputation - Human-centric principle - Physical impact - Psychological impact - Financial impact - Economy - Critical infrastructure - Medical care
Controllability	Unintended output or behavior by administrators	- Integrity	- Adversarial examples [34, 48] - Prompt injection [44] - Indirect prompt injection [13] - Backdoor attack [14, 38] - Cyber attack	- Defensive method for integrity	- Disinformation - Usability - Consumer fairness - Reputation - Human-centric principle - Critical infrastructure - Compliance with laws and regulations - Physical impact - Psychological impact - Financial impact - Economy - Medical care
Trustworthiness	Difficulty in determining the trustworthiness of AI output	- Integrity - Explainability	- Hallucination	- Quantification of uncertainty [58] - RAG [3] - XAI [41] - Detection of hallucination [51]	- Usability - Reputation - Psychological impact - Transparency - Economy - Critical infrastructure - Medical care

society. Most negative impacts on the EIA may not only derive from negative impacts in the ISA, but may also be related to other impacts within the EIA. For example, negative impacts from cyber attacks or disinformation can be the cause of negative economic impacts on non-consumers. A common feature among these elements is that the negative impacts caused by the misuse of AI can occur even when the elements of the ISA are satisfied. These negative impacts do not necessarily affect individuals or society immediately when AI is misused. Ultimately, when the malicious user achieves their specific objective through the AI misuse, negative impacts are inflicted on individuals and society. In the following subsections, we provide detailed interpretation of negative impacts on each security target in the EIA.

Takeaway

There are many negative impacts on individuals and society. The elements related to the misuse of AI do not necessarily have an immediate impact at the moment they are misused. The impact tends to spread when the objective of the misuse is achieved.

4.3 Negative Impacts on Each Security Target

Consumers. Table 3 shows the AI security map for negative impacts on consumers in the EIA. We identified 12 negative impacts on consumers and 10 related elements. It was found that most of the negative impacts on consumers are the result of the compromise of elements in the ISA. It is intuitive that negative impacts on consumers, which is users using AI are associated with negative impacts on information systems or AI itself. On the other hand, some impacts can be caused directly by specific attacks. For example, social engineering attacks [39] and poisoning attacks on RAG [28] can lead to negative impacts related to privacy and misinformation, respectively. Physical, psychological, and financial impacts may occur when human-centered principles are compromised.

Non-consumers. Table 4 shows the AI security map for negative impacts on non-consumers in the EIA. For non-consumers, we identified 13 negative impacts and 10 related elements. Similar to consumers, many of these negative impacts result from the compromise of ISA elements. However, it is important to note that even when ISA elements are functioning properly, there can still be negative impacts if these elements are misused. For example, abusing the accuracy or availability of AI can facilitate the spread of misinformation or enable cyber attacks. In addition, many negative impacts are caused by attacks exploiting AI or by the compromise of other EIA elements. In our map, four negative impacts related to privacy were identified, indicating that there is a high risk of privacy violations for users who do not use AI themselves. This demonstrates the significant external influence that AI can have.

Society. Table 5 shows the AI security map for negative impacts on society and AI system providers in the EIA. We identified nine negative impacts on society and eight related elements. It is found that the compromise of ISA elements is closely linked to impacts on society as well. In particular, impacts on healthcare and critical infrastructure are more likely to occur, due to the large number of relevant ISA elements involved. Given the anticipated widespread utilization of AI in society, it is important to consider potential impacts on these fields and to develop appropriate countermeasures. Furthermore, the misuse of ISA elements can also lead to the spread of disinformation, cyber attacks, and violations of laws and regulations. For example, the dissemination of disinformation through deepfakes is a representative negative impact, resulting from both the availability and the abuse of deepfake technology and its accuracy.

AI system providers. We identified two negative impacts on AI system providers and two related elements. We believe that impacts

Table 3: AI Security Map for negative impacts on consumers in the EIA.

Elements	Negative impacts	Compromised elements in the ISA	Causal factors or related elements in the EIA	Defenses or countermeasures
Privacy	Consumers accidentally inputting their personal information into generative AI or similar systems	- Transparency	- Social engineering attack [39]	- Anonymization technology [47] - DP [1] - Federated learning [52] - Machine unlearning [54] - Encryption technology [18]
Misinformation	Outputting misinformation by AI	- Integrity - Accuracy - Controllability - Explainability - Trustworthiness	- Poisoning attack on RAG [28] - Hallucination	- Defensive method for integrity - Data curation [30] - RAG [3] - XAI [41] - Detection of hallucination [51]
Usability	The decline in the usability of AI	- Integrity - Availability - Accuracy - Controllability - Output fairness		- Defensive methods for integrity and availability - RAG [3]
Consumer fairness	Loss of job or life opportunities due to biased AI output	- Integrity - Controllability - Output fairness		- Defensive method for integrity - Human in the loop - Countermeasures for output fairness - Detection of bias in AI output
	Unfair biased and discriminatory output	- Integrity - Controllability - Output fairness		- Defensive methods for integrity - AI alignment [32] - Countermeasures for output fairness - Detection of bias in AI output
Transparency	Unintentionally using AI	- Explainability		- Identification of AI-generated output - Watermarking for generative AI [20]
	Using AI without recognizing the risks	- Explainability - Trustworthiness		- AI-generated output with disclaimers - Education and follow-up
Human-centric principle	Improperly manipulating the decision-making of consumers by AI	- Integrity - Explainability - Controllability - Output fairness		- Defensive methods for integrity - Human in the loop
Ethics	Unethical output or actions by AI	- Integrity	- Jailbreak	- Education and follow-up - AI alignment [32]
Physical impact	Physical harm to consumers caused by AI	- Integrity - Accuracy - Controllability - Safety	- Human-centric principle	- Defensive methods for integrity
Psychological impact	Psychological harm to consumers caused by AI	- Integrity - Availability - Controllability - Safety - Output fairness	- Consumer fairness	- Defensive methods for integrity - Defensive methods for availability
Financial impact	Financial harm to consumers caused by AI	- Integrity - Availability - Accuracy - Controllability - Safety - Output fairness	- Human-centric principle	- Defensive methods for integrity - Defensive methods for availability

on AI system providers can result from the compromise of any of the ISA elements. Furthermore, negative impacts caused by misuse can lead to reputational and financial consequences for AI system providers. As the number of companies and individuals developing AI-based systems increases, it is important that these risks are properly recognized.

Takeaway

Most negative impacts on **consumers** tend to be linked to the compromise of AI elements in the ISA. However, some can result directly from targeted attacks, such as social engineering [39] or poisoning [4, 5], leading to privacy and misinformation issues for consumers. **Non-consumers** can also experience negative impacts, due to the abuse of AI elements by attackers, especially in the form of privacy violations and the spread of misinformation. **Society** can incur significant impacts, particularly concerning critical infrastructure and medical care. AI misuse can cause widespread harm such as disinformation, cyber attacks, and legal violations. For **AI system providers**, negative impacts can lead to reputation and financial damage, and these risks increase as more entities develop AI systems.

4.4 Relationships between Two Aspects

Basically, we assume that the compromise of any element in the ISA have impacts on the elements of the EIA. In a nutshell, we consider that negative impacts on individuals and society may occur in a chain reaction from the compromise of information systems in many cases. By referring elements categorized into the two aspects, we can reveal how negative impacts affect individuals and society. We found that there are two types of chains leading to negative impacts on individuals and society. The first type is the case where the compromise of AI in information systems has a direct influence on individuals and society. In this case, negative impacts in the ISA are directly related to those on individuals and society. The second type is the case where negative impacts affect individuals and society indirectly, through multiple negative impacts in the EIA. In what follows, we describe each of these chains in detail.

Table 4: AI Security Map for negative impacts on non-consumers in the EIA.

Elements	Negative impacts	Related elements in the ISA		Causal factors or related elements in the EIA	Defenses or countermeasures
		Compromise	Abuse		
Cyber attack	Using AI for cyber attacks	- Confidentiality - Controllability	- Availability - Accuracy - Explainability		- AI alignment [32] - Method for providing explainability while concealing model information
Military use	Using AI for military purposes	- Controllability	- Availability - Accuracy - Explainability		- AI alignment [32]
Privacy	Privacy violation due to the leakage of personal information from AI	- Confidentiality - Integrity			- DP [1] - Federated learning [52] - AI alignment [32] - Machine unlearning [49] - Encryption technology [18] - Anonymization technology [47]
	Using fragmented information and AI to make inferences and identify personal information			- Attacks that use AI to analyze information collected from social media to identify individuals	- Anonymization technology [47] - DP [1] - Federated learning [52] - Machine unlearning [49] - Encryption technology [18]
	Inferring a person's character or personal preferences from their facial expressions			- Attacks that use AI to analyze images to infer personal information	- Anonymization technology [47] - DP [1] - Federated learning [52] - Machine unlearning [49] - Encryption technology [18]
	Using AI-generated emails or audio to prompt the input of confidential information and steal it		- Availability - Accuracy	- Social engineering attack [39]	- Anonymization technology [47] - DP [1] - Federated learning [52] - Machine unlearning [49] - Encryption technology [18]
Disinformation	Creating disinformation using AI	- Controllability - Integrity	- Availability - Accuracy	- Deepfake - Social engineering attack [39]	- AI alignment [32] - Watermarking for generative AI [20] - Encryption technology [18] - Identification of AI-generated output - Detection of disinformation - Deepfake detection [53]
Copyright and authorship	Violation of copyright and authorship by AI-generated similar content	- Integrity - Controllability		- Plagiarism	- Defensive methods for integrity and plagiarism
Human-centric principle	Improperly manipulating the decision-making of non-consumers by AI			- Disinformation	- Defensive methods for integrity and disinformation - Human in the loop
Ethics	Unethical output or actions by AI	- Integrity		- Jailbreak	- Education and follow-up - AI alignment [32]
Physical impact	Physical harm to non-consumers caused by AI	- Integrity - Accuracy - Controllability - Safety		- Military use	- Defensive methods for integrity
Psychological impact	Psychological harm to non-consumers caused by AI	- Confidentiality - Controllability - Safety - Output fairness		- Disinformation - Military use	- Defensive methods for integrity - Defensive methods for availability - Defensive methods for disinformation
Financial impact	Financial harm to non-consumers caused by AI	- Safety		- Cyber attacks - Disinformation	- Defensive methods for integrity - Defensive methods for disinformation

Direct chain of impacts. Negative impacts caused by the compromise of elements within the ISA often directly affect individuals and society. A typical example is privacy. It is easy to imagine that when "confidentiality" is compromised, "privacy" is also breached. In the context of AI security, where AI can retain knowledge about vast amounts of data, the impact on non-consumers can be significant. Therefore, from the perspective of promoting AI utilization, technologies that ensure confidentiality and privacy protection are considered important. Furthermore, as shown in Table 1, negative impacts caused by the integrity violation are related to many external elements, such as human-centered principles, medical care, and critical infrastructure. As a result, it has been found that a compromise of integrity, in particular, can have a significant social impact. As with the integrity violation, negative impacts resulting from the availability breach directly relate to many elements in the EIA.

Indirect chain of impacts. In addition to cases where individuals or society are directly affected, there are also cases where multiple impacts occur in a chain, ultimately affecting individuals or society. For example, as shown in Table 2, if the "integrity" of an LLM is compromised through a prompt injection attack, "controllability" is first affected. Once controllability is compromised, attackers can make the LLM generate "disinformation" as they intend. If this disinformation spreads, it can influence human decision-making, thereby causing negative effects in terms of "human-centered principles" as shown in Table 4. As a result, non-consumers who see this disinformation may also be affected. Furthermore, the misuse of AI systems can occur by taking advantage of the fact that elements in the ISA are satisfied. This can also have a cascading effect on individuals and society. For instance, cyber attacks through malware generation using LLMs exploit "accuracy" and "availability", leading to "financial impacts" on non-consumers and negative "economic" effects on society. In particular, such ripple effects tend to influence non-consumers who do not use AI at all. It is important to consider how to prevent negative impacts on non-consumers

Table 5: AI Security Map for negative impacts on society and AI system providers in the EIA.

Elements	**Negative impacts**	**Related elements in the ISA**		**Causal factors or related elements in the EIA**	**Defensive methods or countermeasures**
		Compromise	**Abuse**		
		Society			
Cyber attack	Using AI for cyber attacks	- Confidentiality - Controllability	- Availability - Accuracy - explainability		- AI alignment [32] - Method for providing explainability while concealing model information
Military use	Using AI for military purposes	- Controllability	- Availability - Accuracy - Explainability		- AI alignment [32]
Disinformation	Creating disinformation using AI	- Controllability	- Availability - Accuracy	- Deepfake - Social engineering attack [39]	- AI alignment [32] - Watermarking for generative AI [20] - Encryption technology [18] - Identification of AI-generated output - Detection of disinformation - Deepfake detection [53]
Compliance with laws and regulations	Using AI for purposes that violate the law	- Confidentiality - Controllability	- Availability - Accuracy		- AI alignment [32] - AI access control
	Actions that violate the law by AI	- Integrity - Accuracy - Controllability			- AI alignment [32] - AI access control
Ethics	Unethical output or actions by AI	- Integrity		- Jailbreak	- Education and follow-up - AI alignment [32]
Economy	AI negatively impacting the economy	- Safety - Accuracy		- Cyber attack - Military use - Disinformation - Human-centric principle	- Defensive methods for integrity - Defensive methods for disinformation
Medical care	Negative impact on medical care caused by AI	- Integrity - Availability - Accuracy - Controllability - Safety - Output fairness - Explainability - Trustworthiness		- Human-centric principle	- Defensive methods for integrity - Defensive methods for availability
Critical infrastructure	Negative impact on critical infrastructure caused by AI	- Integrity - Availability - Accuracy - Controllability - Safety - Output fairness - Explainability - Trustworthiness		- Human-centric principle	- Defensive methods for integrity - Defensive methods for availability
		AI system providers			
Reputation	The decline in the reputation of AI system providers	- All the elements in the ISA		- Cyber attack - Military use - Compliance with laws and regulations	- Defensive methods for confidentiality - Defensive methods for integrity - Defensive methods for availability
Financial impact	Financial harm to AI system providers caused by AI	- All the elements in the ISA		- Cyber attack - Compliance with laws and regulations	- Defensive methods for integrity - Defensive methods for availability

and society from such misuse, while also ensuring convenience for consumers.

Takeaway

There are the two types of chains in terms of impacts on people and society. In particular, the misuse of AI can occur by exploiting the fulfilled elements of the ISA. Therefore, in addition to AI-specific defense methods, considering alternative defense approaches with zero trust security helps mitigate negative impacts on individuals and society.

5 Recommendations

We present some recommendations in response to the current landscape of AI security and the holistic organization in our work. We consider recommendations from the two perspectives, namely fundamental research and applied research.

5.1 Fundamental Research on AI Security

As shown above, the impacts of AI and ML extend beyond traditional security concepts and the scope of privacy. Hence, we encourage researchers working on future studies to holistically analyze what kinds of impacts may occur with holistic overviews such as our map when they consider new attacks or countermeasures. By doing this, it may help them devise more innovative research directions and more practical methods, thereby enhancing AI security. We hope that our AI security map will contribute to the promotion of fundamental research, as well as the discovery of new elements, negative impacts, and security targets.

Furthermore, we recommend that more active efforts be made to pursue holistic organization as we have proposed. In what follows, we present several **open problems** and future directions in terms of the holistic organization of AI security for researchers.

Appropriate Granularity in Defining security targets. At this stage, security targets in our map are defined in relatively broad categories. However, in reality, a more diverse range of stakeholders should be considered. For example, since “consumers” may include both decision-makers and developers, each of whom may require different types of information. Therefore, it is necessary to establish a more detailed definition of stakeholders in future work.

Quantitative Assessment of Impacts. Our map helps understand a holistic overview of AI security. However, the degree of each impact is not clear at this stage. Currently, only the relationships among various factors are identified, and quantitative risk assessment has not yet been addressed. To enhance the informativeness of the analysis with the AI security map, it is important to indicate the extent of risk associated with each negative impact in future. Since the level of risk may vary depending on the security targets and the chain of impacts, it is important to develop quantification methods that take these factors into account.

Mapping New Domains. It is also necessary to consider new domains such as the security of agentic AI. Additionally, there is room for discussion regarding the positioning of methods such as safety verification and red teaming, which are not direct countermeasures. While these techniques are important for proactively identifying vulnerabilities and assessing risks in AI systems, their classification should be carefully considered, as they are more akin to preventive measures. Due to the extremely rapid emergence of new AI technologies, it is also necessary to establish mechanisms and methodologies for efficiently keeping up with these advancements and promptly reflecting them in the holistic organization.

Automation of the Definition and Classification of Negative Impacts and Elements. At this stage, elements and negative impacts are identified and classified manually by referencing multiple papers and guidelines. Given the rapid emergence of new AI technologies, it may become necessary to automatically define and classify elements in response to the swift evolution of AI. This challenge is also relevant to conventional survey papers because it is likely to be difficult to comprehensively cover the vast amount of information manually without omissions. Thus, the development of systematization technologies utilizing LLMs and AI is interesting and essential in future.

5.2 Applied Research on AI Security

In addition to fundamental research, it is also important to consider how the holistic overview can be utilized and further developed from the perspective of applied research. For people other than AI security researchers and experts, it is important to be aware of the potential attacks and risks associated with the utilization of AI. In particular, business stakeholders and decision-makers who consider incorporating AI into their services or products may be interested in negative impacts entailed by introducing AI. For example, one possible application is obtaining relevant negative impacts by inputting a news article about AI security into an LLM with our map. By doing this, they can gain insights into relevant negative impacts without expertise. We actually conducted a basic trial of this use case. To this end, we input a news article about AI being misused for cyber attacks into GPT-4.1 [31]. The prompt used for this use case is shown in Appendix C. As a result, the LLM outputs negative impacts regarding four elements (integrity, availability, trustworthiness, and accuracy) in the ISA and two elements (cyber attacks and human-centric principle) in the EIA. While elements related to cyber attacks can be easily inferred from the content of the news, "human-centric principle" is not as readily apparent. Therefore, this result demonstrates that valuable information in the EIA can be obtained. Such a use case is also expected to be highly beneficial for consumers.

Another application of our map is that AI system providers or developers use our map in order to to understand negative impacts on AI within information systems. It is essential for them to be aware of the potential negative impacts in the EIA when designing and implementing information systems. Depending on the service, the primary concern required to minimize negative impacts on end users is different. Thus, such a use case is helpful in determining which element in the ISA should be developed with greater robustness. Being informed of these possible impacts in advance enables AI system providers or developers to anticipate adverse outcomes and design systems more efficiently and effectively.

By implementing the above use cases, it is possible to facilitate a better understanding of the complex technologies involved in AI security, the potential negative impacts on each security target, and the relationships among various elements. Since the above use cases are just examples, other useful application or practical methods should be devised in future.

6 Conclusion

In this paper, we have developed the AI security map to provide a holistic overview of AI security. The map identifies key elements that AI systems should fulfill, as well as factors that influence individuals and society, from both the perspective of the ISA and EIA. Furthermore, we holistically organized the relationships between AI elements and external factors in the EIA, which had not previously been addressed. In short, our map helps clarify how damage to AI-based information systems can affect people and society. We also categorized the specific negative impacts associated with each element, along with their causes including attacks, causal factors, and countermeasures. By distinguishing whether these causes originate from the compromise or misuse of elements in the ISA, we clarify how AI can lead to various consequences. Through the insights and recommendations derived from our map, we discuss the value of holistically organizing AI security knowledge and technology. We hope that our work will serve as an important foundation for researchers and a wide range of stakeholders, facilitating the collection and understanding of AI security information in this complex field, and promoting further discovery and research on new elements and negative impact targets.

Acknowledgments

This work was supported by JST K Program, Japan Grant Number JPMJKP24C4.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining* 13, 1 (2023), 30.

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389* (2012).

[5] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. 2024. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 407–425.

[6] Alberto Castagnaro, Mauro Conti, and Luca Pajola. 2024. Offensive AI: Enhancing Directory Brute-forcing Attack with the Use of Language Models. In *Proceedings of the 2024 Workshop on Artificial Intelligence and Security*. 184–195.

[7] Yuanhaur Chang, Han Liu, Evin Jaff, Chenyang Lu, and Ning Zhang. 2024. SoK: Security and Privacy Risks of Medical AI. *arXiv preprint arXiv:2409.07415* (2024).

[8] Huaming Chen and M Ali Babar. 2024. Security for machine learning-based software systems: A survey of threats, practices, and challenges. *Comput. Surveys* 56, 6 (2024), 1–38.

[9] Sayanton V Dibbo. 2023. Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. IEEE, 439–456.

[10] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. 2020. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proceedings of the 36th Annual Computer Security Applications Conference*. 897–912.

[11] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33.

[12] Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. 2024. Privacy and security concerns in generative AI: a comprehensive survey. *IEEE Access* (2024).

[13] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 79–90.

[14] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733* (2017).

[15] Yupeng Hu, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. 2021. Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–36.

[16] Zhanhao Hu, Julien Piet, Geng Zhao, Jiantao Jiao, and David Wagner. 2024. Toxicity detection for free. *arXiv preprint arXiv:2405.18822* (2024).

[17] Bo Hui, Haolin Yuan, Neil Gong, Philippe Burlina, and Yinzhi Cao. 2024. Pleak: Prompt leaking attacks against large language model applications. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 3600–3614.

[18] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. 2018. {GAZELLE}: A low latency framework for secure neural network inference. In *27th USENIX security symposium (USENIX security 18)*. 1651–1669.

[19] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. 2024. The pursuit of fairness in artificial intelligence models: A survey. *arXiv preprint arXiv:2403.17333* (2024).

[20] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning*. PMLR, 17061–17084.

[21] Naoto Kiribuchi, Kengo Zenitani, and Takayuki Semitsu. 2025. Securing AI Systems: A Guide to Known Attacks and Impacts. arXiv:2506.23296 [cs.CR] https://arxiv.org/abs/2506.23296

[22] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[23] Linyi Li, Tao Xie, and Bo Li. 2023. Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)*. IEEE, 1289–1310.

[24] Lin Lu, Hai Yan, Zenghui Yuan, Jiawen Shi, Wenqi Wei, Pin-Yu Chen, and Pan Zhou. 2024. Autojailbreak: Exploring jailbreak attacks and defenses through a dependency lens. *arXiv preprint arXiv:2406.03805* (2024).

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[26] Melkamu Mersha, Khang Lam, Joseph Wood, Ali AlShami, and Jugal Kalita. 2024. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* (2024), 128111.

[27] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. 2021. A survey on security and privacy of federated learning. *Future Generation Computer Systems* 115 (2021), 619–640.

[28] Fatemeh Nazary, Yashar Deldjoo, and Tommaso di Noia. 2025. Poison-rag: Adversarial data poisoning attacks on retrieval-augmented generation in recommender systems. In *European Conference on Information Retrieval*. Springer, 239–251.

[29] Maximilian Noppel and Christian Wressnegger. 2024. SoK: Explainable machine learning in adversarial environments. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2441–2459.

[30] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749* (2021).

[31] OpenAI. 2025. *Introducing GPT-4.1 in the API*. https://openai.com/index/gpt-4-1

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[33] Rahul Pankajakshan, Sumitra Biswal, Yuvaraj Govindarajulu, and Gilad Gressel. 2024. Mapping llm security landscapes: A comprehensive stakeholder risk assessment proposal. *arXiv preprint arXiv:2403.13309* (2024).

[34] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 372–387.

[35] Otavio Parraga, Martin D More, Christian M Oliveira, Nathan S Gavenski, Lucas S Kupssinskü, Adilson Medronha, Luis V Moura, Gabriel S Simões, and Rodrigo C Barros. 2025. Fairness in Deep Learning: A survey on vision and language research. *Comput. Surveys* 57, 6 (2025), 1–40.

[36] Miguel A Ramirez, Song-Kyoo Kim, Hussam Al Hamadi, Ernesto Damiani, Young-Ji Byon, Tae-Yeon Kim, Chung-Suk Cho, and Chan Yeob Yeun. 2022. Poisoning attacks and defenses on artificial intelligence: A survey. *arXiv preprint arXiv:2202.10276* (2022).

[37] Maria Rigaki and Sebastian Garcia. 2023. A survey of privacy attacks in machine learning. *Comput. Surveys* 56, 4 (2023), 1–34.

[38] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. 2020. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11957–11965.

[39] Marc Schmitt and Ivan Flechais. 2024. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review* 57, 12 (2024), 1–23.

[40] Saskia Laura Schröer, Giovanni Apruzzese, Soheil Human, Pavel Laskov, Hyrum S Anderson, Edward WN Bernroider, Aurore Fass, Ben Nassi, Vera Rimmer, Fabio Roli, et al. 2025. SoK: On the offensive potential of AI. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 247–280.

[41] Gesina Schwalbe and Bettina Finzel. 2024. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* 38, 5 (2024), 3043–3101.

[42] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).

[43] Junjie Shen, Ningfei Wang, Ziwen Wan, Yunpeng Luo, Takami Sato, Zhisheng Hu, Xinyang Zhang, Shengjian Guo, Zhenyu Zhong, Kang Li, et al. 2022. Sok: On the semantic ai security in autonomous driving. *arXiv preprint arXiv:2203.05314* (2022).

[44] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 1671–1685.

[45] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[46] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv preprint arXiv:2408.12622* (2024).

[47] Djordje Slijepčević, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer. 2021. k-anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Computers & Security* 111 (2021), 102488.

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[49] Florian Tramer. 2022. Detecting adversarial examples is (nearly) as hard as classifying them. In *International conference on machine learning*. PMLR, 21692–21702.

[50] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*. 601–618.

[51] Simon Valentin, Jinmiao Fu, Gianluca Detommaso, Shaoyuan Xu, Giovanni Zappella, and Bryan Wang. 2024. Cost-effective hallucination detection for llms. *arXiv preprint arXiv:2407.21424* (2024).

[52] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. 2020. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440* (2020).

[53] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In

Table 6: Elements in the ISA

Elements that AI should satisfy	Definition
Confidentiality	AI data and models are not accessed by unauthorized individuals.
Integrity	The AI models and algorithms have not been tampered with, and the AI outputs are as expected.
Availability	AI can provide the necessary features and services when needed.
Explainability	AI can explain the basis and process of its output.
Output Fairness	AI does not produce biased outputs towards specific individuals or groups.
Safety	AI is equipped with safety mechanisms to prevent harm to human life, body, property, or mind.
Accuracy	AI meets a certain level of accuracy for achieving objectives.
Controllability	AI is controlled by administrators and does not run amok or affect other environments.
Trustworthiness	Output from AI is reliable.

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 8695–8704.

[54] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. 2021. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577* (2021).

[55] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency.* 214–229.

[56] Tatjana Wingarz, Anne Lauscher, Janick Edinger, Dominik Kaaser, Stefan Schulte, and Mathias Fischer. 2024. SoK: Towards Security and Safety of Edge AI. *arXiv preprint arXiv:2410.05349* (2024).

[57] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. 2021. Detecting ai trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP).* IEEE, 103–120.

[58] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems* 37 (2024), 15356–15385.

[59] Qingzhao Zhang, Ziyang Xiong, and Z Morley Mao. 2024. Safeguard is a Double-edged Sword: Denial-of-service Attack on Large Language Models. *arXiv preprint arXiv:2410.02916* (2024).

[60] Shaobo Zhang, Yimeng Pan, Qin Liu, Zheng Yan, Kim-Kwang Raymond Choo, and Guojun Wang. 2024. Backdoor attacks and defenses targeting multi-domain ai models: A comprehensive review. *Comput. Surveys* 57, 4 (2024), 1–35.

[61] Wenbin Zhang. 2024. AI fairness in practice: Paradigm, challenges, and prospects. *Ai Magazine* 45, 3 (2024), 386–395.

[62] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 253–261.

A Definition of Elements in Two Aspects

Table 6 shows definitions of elements in the ISA. Table 7 shows definitions of elements in the EIA.

B Definition of security targets in our work

Table 8 shows definition of security targets in our work.

C Prompt example

The following example is a prompt used for the trial in Section 5.

Table 7: Elements in the EIA

Elements that impact individuals and society	Definition
Cyber attack	AI is not used for cyber attacks.
Military use	AI is not used for military purposes.
Privacy	AI does not infringe on privacy and complies with privacy laws and customs.
Disinformation	AI is not used to intentionally create disinformation, or it can identify such disinformation.
Misinformation	AI does not output misinformation, or it can identify such misinformation.
Usability	AI meets a certain level of usability to achieve objectives.
Consumer fairness	No harm is caused by unfair biased output from AI.
Plagiarism	AI is not used for plagiarism.
Copyright and authorship	AI complies with laws and customs related to copyright and authorship.
Transparency	It is clearly stated that the system uses AI, including information on its limitations and risks associated with its use.
Reputation	The AI system provider is evaluated to a certain standard and is trusted.
Compliance with laws and regulations	AI is used for lawful purposes and produces output or actions that comply with the law.
Human-centric principle	AI is appropriately used for the benefit of humans.
Ethics	AI behaves in a manner consistent with societal norms.
Economy	The use of AI has a positive impact on economy.
Physical impact	The use of AI does not cause physical harm to people.
Psychological impact	The use of AI does not cause psychological harm to people.
Financial impact	The use of AI does not cause financial harm to people.
Medical care	The use of AI contributes to the development of advanced and safe medical care.
Critical infrastructure	The use of AI contributes to the safe operation of critical infrastructure.

Table 8: Security targets in AI security map

Security target	Definition
Consumer	An individual or organization that utilizes AI or AI systems.
Non-consumer	An individual or organization that is not classified as a consumer.
Society	A group composed of multiple people and organizations.
AI system provider	An individual or organization that provides an information system using AI (AI system).

Prompt example

```
Analyze the following news article and extract the negative impacts of AI technology on individuals, organizations, or society that are explicitly stated in the article based on the specified list of negative impacts. If a negative impact is not yet apparent and there is only a possibility the negative impact may occur in the future, it should be excluded. The extracted negative impacts should be output in JSON format.
# List of Negative Impacts:
- Decrease in AI prediction accuracy
- Personal information leakage
- Using AI for cyber attacks
....
# News Article
[input news here]

# Output Format:
{
"Negative impacts": [
{ "impact": "[Negative impact]",
"description": "[Explanation or relevant part of the news article]" },
// Add other negative impacts as needed
] }
```