SoK: How Artificial-Intelligence Incidents Can Jeopardize Safety and Security

Richard May
Harz University of Applied Sciences
Wernigerode, Germany
rmay@hs-harz.de

Jacob Krüger
Eindhoven University of Technology
Eindhoven, The Netherlands
j.kruger@tue.nl

Thomas Leich
Harz University of Applied Sciences
Wernigerode, Germany
tleich@hs-harz.de

ABSTRACT

In the past years, a growing number of highly-automated systems has build on Artificial-Intelligence (AI) capabilities, for example, self-driving vehicles or predictive health-state diagnoses. As for any software system, there is a risk that misbehavior occurs (e.g., system failure due to bugs) or that malicious actors aim to misuse the system (e.g., generating attack scripts), which can lead to safety and security incidents. While software safety and security incidents have been studied in the past, we are not aware of research focusing on the specifics of AI incidents. With this paper, we aim to shed light on this gap through a case survey of 240 incidents that we elicited from four datasets comprising safety and security incidents involving AI from 2014 to 2023. Using manual data analyses and automated topic modeling, we derived relevant topics as well as the major issues and contexts in which the incidents occurred. We find that the topic of AI incidents is, not surprisingly, becoming more and more relevant, particularly in the contexts of autonomous driving and process-automation robotics. Regarding security and its intersection with safety, most incidents connect to generative AI (i.e., large-language models, deep fakes) and computer-vision systems (i.e., facial recognition). This emphasizes the importance of security to also ensure safety in the context of AI systems, with our results further revealing a high number of serious consequences (system compromise, human injuries) and major violations of confidentiality, integrity, availability, as well as authorization. We hope to support practitioners and researchers in understanding major safety and security issues to support the development of more secure, safe, and trustworthy AI systems.

CCS CONCEPTS

• **Security and privacy**; • **Hardware → Safety critical systems**; • **Computing methodologies → Artificial intelligence**;

KEYWORDS

safety, security, safety-critical systems, vulnerabilities, artificial intelligence, machine learning

ARES 2024, July 30-August 2, 2024, Vienna, Austria

ACM ISBN 979-8-4007-1718-5/24/07…$15.00
https://doi.org/10.1145/3664476.3664510

ACM Reference Format:
Richard May, Jacob Krüger, and Thomas Leich. 2024. SoK: How Artificial-Intelligence Incidents Can Jeopardize Safety and Security. In *The 19th International Conference on Availability, Reliability and Security (ARES 2024), July 30-August 2, 2024, Vienna, Austria.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3664476.3664510

1 INTRODUCTION

The amount of highly-automated systems that build on Artificial Intelligence (AI) has rapidly increased in recent years [20, 46]. Such systems typically use complex machine-learning functionalities aimed at solving real-world problems in a variety of domains, such as automotive, manufacturing, healthcare, or education [35, 61]. These AI systems enable various novel or enhance existing functions, such as autonomous driving [42], automated and predictive manufacturing [52], automated medical diagnostics [20], or advanced natural-language processing [65]. AI systems offer great opportunities, particularly to optimize flexibility, scalability, and efficiency, often resulting in novel business models as well as tremendous time or cost savings [3, 52].

However, using AI systems in high-stakes contexts entails critical risks that arise from system behavior that may not correspond to the intended one (i.e., misbehavior) or from intentionally wrong use (i.e., misuse) [41, 55]. For such reasons, AI systems are widely known as risky actors that can promote unintended events and consequent harms to the environment, systems, or even humans [73, 82]. Such events are typically referred to as *AI incidents* [41, 55]. AI incidents are quite diverse, including, for example, failures in the context of self-driving vehicles, inadequate labeling [10, 25], or AI-powered networks [39]. Overall, AI systems can often pose considerable risks to safety and security, involving, for example, issues related to system vulnerabilities that can be exploited to manipulate the AI and its models [44], malicious control of driver-assistance systems through over-the-air updates [84], or misdiagnoses in medical decision-support systems [16]. Thus, ensuring safety and security in AI systems has become essential, particularly if misbehavior or misuse can jeopardize human lives (i.e., safety-critical systems). Ensuring system safety and security is already highly complex due to trends like increasing system configurability [2, 37, 50, 51] and the growing number of security attacks [45, 86], but AI systems involve novel risks with unknown impact—including risks related to the AI's own complexity.

The misbehavior or misuse of AI systems is commonly discussed in controversial ways [73], leading to various reports on such issues. To provide comprehensive overviews of AI incidents, datasets listing thousands of entries have been created recently, for example, the AI Incident Database (AIID). Such datasets are typically aimed at offering information to learn from past problems as well

as to empower criticism and novel solutions to tackle these problems [55, 67, 73]. Accordingly, these datasets are a valuable source for analyzing AI incidents, for instance, to understand when, how, and why they occurred, as well as whether and to what extent they pose risks or caused harm.

In this paper, we present a case survey based on reports from four AI-incident datasets, namely the *AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC)* repository, the *AIID*, the *AI Vulnerability Database (AVID)*, and *Where in the World is AI? (WitWiAI)*. We investigated 240 AI incidents from the past decade (2014–2023) that posed risks and violated safety and security. To the best of our knowledge, this case survey presents the first analysis of these AI incident datasets in the context of safety and security. Moreover, we are not aware of comparable works that triangulated from these four established datasets on AI incidents. With our work, we aim to explore current trends regarding AI incidents with respect to safety and security, their characteristics, and their contexts. This way, our research goal is to **provide an understanding of what and how AI incidents relate to safety and security**.

More specifically, we contribute in this paper:

- An overview of the main safety and security issues as well as trends regarding AI incidents.
- A discussion of the incidents, topics, and technologies regarding the connections between AI, safety, and security.
- An open-access dataset (i.e., the analysis file) including the 240 harmonized entries on safety and security that we extracted from the four datasets.[1]

Our insights on AI incidents offer a new, valuable, and practice-relevant perspective on the context of safety and security of AI systems. We aim to support both practitioners and researchers in understanding real-world safety and security issues of AI systems; raising the awareness fore AI incidents and support the development as well as research of more secure, safe, and trustworthy AI systems.

2 BACKGROUND

Next, we provide background on AI systems, the current state of AI incidents datasets, as well as safety and security.

2.1 AI Systems

AI typically refers to the artificial acquisition and application of information based on machine-learning capabilities [26, 70]. These capabilities implement a variety of functions aimed at solving real-world problems (e.g., autonomous driving [42], facial detection [76], predictive maintenance [52]) using a diverse bundle of technologies with various configuration options [52, 70, 71].

AI systems can be roughly classified from a technological perspective according to their learning strategies, tasks and operations, working areas, as well as added value [17, 70, 71]. Learning strategies refer to how information is generated based on the input data, for example, supervised or reinforcement learning [26, 83]. Tasks describe the way AI systems analyze and identify patterns, and the associated operations specify the operationalization of these tasks (i.e., the method used to solve a given problem) [70]. For instance, artificial neural networks (i.e., the operation) can be used for data classification purposes (i.e., the task) [26, 70, 71]. Moreover, operations provide internal model structures (e.g., hyperparameters, activation and loss function) configured for a specific use case [26]. The working areas comprise diverse fields in which the system is used, for example, computer vision, robotics, or generative systems [4, 38, 70]. By applying an AI system in the context of a specific area, a certain added value is typically expected as an economic incentive. These incentives are quite diverse and case-specific, ranging from improving quality and efficiency of processes to reducing time and costs during system deployment or operation [49, 59, 83].

2.2 AI Incidents Datasets

Using AI systems in high-stakes contexts can pose risks that may cause harm, for instance, security breaches may lead to unauthorized data access [21, 82]. If such negative events occur, they are referred to as *AI incidents* [41, 55]. AI incidents are typically triggered by misbehavior (e.g., manufacturing-robot accidents, misconfiguration) or misuse (e.g., malicious deepfakes, attack-script generation) and can range from technical issues (e.g., safety- or security-critical system failure) to ethical or societal concerns (e.g., privacy-related data theft) [25, 61, 79, 82].

To document information on AI incidents in a structured way, several open-access datasets have been created. These comprise meta data (e.g., occurrence year, domain, technology) on incidents that have occurred around the world, typically including various media reports (e.g., newspaper articles, technical reports of companies) [55, 67, 82]. Such datasets aim to help understand the complex behavior of AI systems and why they fail. Since incidents usually occur after a system is deployed [29], AI incidents datasets motivate practitioners, regulators, and researchers to learn from the data, ideally helping to identify and reduce incidents early on during a system's development. Currently, there are four widely established and independent datasets, namely the AIAAIC repository, the AIID, the AVID, and WitWiAI, each of which comprises hundreds of entries on AI incidents [21, 67, 82]. However, it is usually not clear who maintains such a dataset, in particular to protect the identities of these people. Nevertheless, although this fact poses several threats to the validity when analyzing these datasets, especially the AIAAIC repository and the AIID have already been accepted by researchers due to their detailed qualitative entries [21, 22, 55], which have served as the basis for several scientific studies (cf. Section 7).

2.3 Safety and Security

Safety and security are essential quality attributes of any system, especially if it involves weaknesses (e.g., vulnerabilities) whose exploitation could cause harm [31, 32, 37, 62]. Safety (i.e., human safety, functional safety) is aimed at implementing a system so that it causes no harm to itself, its users, or the environment by ensuring that the system operates as intended [31, 60]. Typically, ensuring safety is oriented towards two categories of failures: systematic or hardware failure. These categories are usually addressed by referring to the software integrity level (SIL), which is the fundamental measurement for estimating safety [6, 31].

Security refers to the protection of a software system against unauthorized parties, for example, malicious data access or manipulation, by addressing security goals [19, 32, 62]. These goals

[1] https://doi.org/10.5281/zenodo.11946279

include the well-known CIA triad (i.e., confidentiality, integrity, availability) and three additional goals for information security (i.e., accountability, authorization, non-repudiation) [47, 69]. Overall, security is typically characterized by assets (e.g., threatened data), threats (i.e., unwanted events that may cause harm), risks (i.e., actual exploitation of weaknesses), and countermeasures (e.g., authentication mechanisms) [32, 58, 62].

Although both safety and security pursue similar goals at different levels (i.e., protection), there are various cases where both connect and depend on each other. In particular, security plays an essential role in protecting safety-critical systems [18, 28], for example, cyber-physical systems like predictive-manufacturing systems [52] or vehicles with autonomous driving capabilities [84]. Consequently, security issues typically have great impact on the actual safety of safety-critical systems [18, 28, 51].

3 METHODOLOGY

Next, we detail our methodology, including our research goal and research questions, study design, as well as study conduct.

3.1 Goal and Research Questions

For our case survey, we defined our research goal as providing an understanding of what and how reported AI incidents relate to safety and security. To achieve this goal, we defined two specific Research Questions (RQs):

RQ$_1$ How prevalent are AI incidents that jeopardize safety and/or security?
First, we aimed to assess how AI incidents in the context of safety and security evolved. Precisely, our objective was to provide a chronological overview to analyze the occurrence of these topics over time to reveal trends.

RQ$_2$ How do AI incidents relate to safety and/or security?
Second, we aimed to shed light on the contexts in which the AI incidents occurred, aiming to identify and classify recurring patterns of safety and security issues. By going into the details of these AI incidents and their consequences, we discuss the relations between AI, safety, and security.

Through our study, we aim to provide a comprehensive overview and understanding of the connections between AI incidents, safety, and security. Thereby, we contribute insights for practitioners and researchers that can help design as well as implement more secure, safe, and trustworthy AI systems.

3.2 Study Design

We conducted a case survey by building on published reports that we mined from AI incident databases. Overall, our study design follows common recommendations for mining datasets in software engineering [85] and we aimed to meet the respective data quality criteria during the evaluation process as well as the subsequent interpretation of the data [8]. Our actual analysis of the cases' AI incidents is based on two methods (cf. Figure 1), namely (1) *a manual data analysis* and (2) *an automated topic modeling* based on Latent Dirichlet Allocation (LDA) [7].

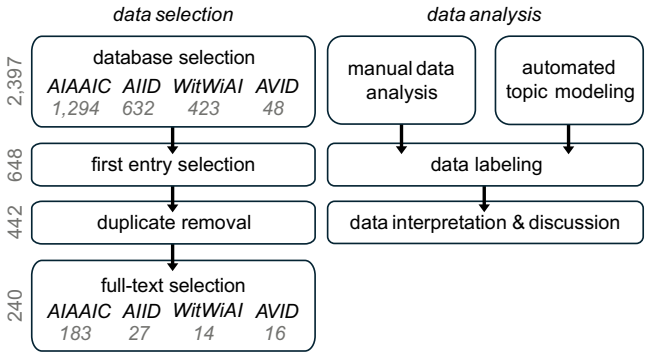

Figure 1: Methodological overview of our study (i.e., data selection and analysis). The gray numbers indicate the total number of AI incidents we considered after each step.

Datasets. We relied on four AI incident databases, namely the *AIAAIC* repository,[2] the *AIID*,[3] the *AVID*,[4] and *WitWiAI*.[5] The AIID, the AVID, and WitWiAI are largely involving entries that are listed in the AIAAIC repository, too—but they also include a number of different entries, which helps us to gain even broader insights into the topic. Specifically, while the AIAAIC and AIID focus on generally collecting and cataloging AI incidents, the WitWiAI lists incidents based on the geographical distribution of incidents. In addition, the AVID includes only incidents related to security-related weaknesses. In the context of our study, we relied on the categories of the AIAAIC repository and transferred them to all datasets, as researchers consider them to be reasonably structured [67, 82]. These categories are represented via 16 open-access data entries, for instance, on an incident's occurrence years, domains, systems, risks, and an URL. Through the URL, it is possible to access a corresponding incident entry on the respective website. We argue that the four datasets are highly valuable and appropriate to meet our research goal, because they are currently the largest datasets on this topic with the highest data quality [48, 67, 82], strengthening the reliability of our survey. Identically, that these datasets have been used in published studies, for example, in the context of privacy [15] or domain issues [9], increases our confidence in the datasets.

Data Selection. We defined four Selection Criteria (SCs) to include only entries that are feasible for answering our research questions:

- SC$_1$ The incident occurred during the past decade (2014–2023).
- SC$_2$ The incident sources (i.e., reports) are still available.
- SC$_3$ The entry describes an actual incident (i.e., not an issue that may lead to an incident).
- SC$_4$ The incident poses risks to safety, security, or both.

We restricted the time frame to consider only recent AI incidents, arguing that these entries are more relevant to our study and its goal of understanding temporary issues. Moreover, in recent years, there have been significant developments in the context of AI systems [23], highlighting that older incidents may already be outdated

[2] https://www.aiaaic.org/aiaaic-repository, accessed January 02, 2024.
[3] https://incidentdatabase.ai, accessed January 02, 2024.
[4] https://avidml.org, accessed January 02, 2024.
[5] https://map.ai-global.org, accessed January 02, 2024.

(SC_1). By checking whether the original reports still exist, we intended to ensure that the AI incidents did actually occur and that we could collect as well as verify background information. Note that we cannot fully ensure that each incident actually took place, since a validation of each individual report is out of our scope. Instead, we rely on the quality controls performed by the dataset owners (SC_2). Moreover, the datasets contain both incidents (i.e., an event) and issues (i.e., public concerns on potential impacts) regarding AI systems. We focus solely on incident entries because these are events that actually occurred and do not rely on assumptions (SC_3). Lastly, we ensured that the selected entries are actually part of our study's thematic scope (SC_4).

Extracted Variables. For our manual data analysis, we relied on eight categories from the AIAAIC dataset. Furthermore, we added two more categories as additional classifications of the incidents based on the given data entries (i.e., *context* and *serious consequences*, both named by the authors). We classified each category according to three superordinate topics. Please note that we partly renamed the following categories compraed to the AIAAIC for better understanding (e.g., *sector* into *domain*):

Entry attributes provide the context of an incident, specifically a clear identification as well as where (i.e., domain) and when the incident occurred, including:
- **ID**: The unique identifier of any entry, taken unchanged from the datasets (e.g., AIID-000).
- **Occurrence year**: The year in which the incident occurred for the first time.
- **Domain**: The primary domain of the impacted system (e.g., automotive).

Systems specify in what systems with what specifications the incident occurred, including:
- **Name**: The name of a system that is primarily involved in the incident.
- **Technology**: The technology that was deployed in the system (e.g., computer vision, robotics).
- **Purpose**: The objective goal of the system (e.g., identifying persons).

Incidents characterize each incident, its properties, and its general severity, including:
- **Origin**: A classification of each incident as either *misuse* (i.e., an issue caused by misusing an AI system either technically or ethically) or as *misbehavior* (i.e., an unwanted issue of an AI system that does not correspond to the intended behavior).
- **Risks**: Any additional risks occurring due to the incident (e.g., for reliability, robustness).
- **Serious consequences**: A specification whether an issue was actually exploited (i.e., related to security), resulted in a serious accident (i.e., related to safety), or did not result in serious consequences despite the incident itself.
- **Harms**: The negative impact caused by the incident (e.g., physical injury, system update).

This categorization is feasible for addressing our research questions and defines detailed background on each AI incident.

In addition, we performed a topic modeling according to the guidelines of Agrawal et al. [1] to classify the *working areas* of the systems involved (e.g., prediction systems, generative AI). By doing this, we aimed to provide an additional classification to address **RQ_2**. The topic modeling we used is based on LDA, which is a statistical model that helps uncover topics within a collection of documents (i.e., AI incidents) [7]. For our analysis, we assumed that each AI incident (i.e., title and description taken from the datasets) represents a topic that is characterized by a specific distribution of words. We argue that LDA is appropriate for our study, as it is a well-established method in research related to software engineering [11], including the fields of safety [34] and security [75].

3.3 Study Conduct

In the following, we describe our data selection and analysis process according to our research methodology (cf. Figure 1).

Manual Data Analysis. The first author accessed and downloaded all datasets on January 02, 2024, and merged them in a central Excel spreadsheet. At this point, all datasets together comprised approximately 2,400 AI incidents (AIAAIC repository: 1,294; AIID: 632; WitWiAI: 423; ACID: 48). First, we deleted all incidents that were out of scope according to all our selection criteria (i.e., considering event date, report availability, actual incidents, and safety or security context). This exclusion step resulted in a total of 648 incidents that posed risks to safety and security. Second, the first author deleted duplicates, after which we ended up with 442 entries. Third, we examined the remaining entries in detail based on our selection criteria, which resulted in us discarding 202 entries as not relevant according to our thematic focus and the entries' limited connection to safety and security. Consequently, we considered a total of 240 incidents as relevant with respect to our research goal. Note that we partly relabeled or grouped existing classifications of categories if these were incomplete or difficult to understand from our perspective. All selected incidents were manually validated by the first author to ensure that the classifications are correct and understandable. Lastly, the first author analyzed the 240 AI incidents based on our defined categories to address our research questions.

Topic Modeling. The first author implemented the topic modeling in a Python script, which uses the libraries `NLTK`, `stop_words`, and `gensim`. At first, we reduced the complexity of the AI incidents (i.e., titles and descriptions) by applying four pre-processing steps:

(1) automated hyperlink and tag removal,
(2) tokenization of the texts (i.e., into words),
(3) lemmatization of the tokens (based on Penn Treebank tagset),
(4) vectorization of the classified tokens (based on term frequency-inverse document frequency).

Next, we applied LDA on the pre-processed data, using experimentation to identify the best fit. Specifically, the first author varied the number of topics k from five to 20 with between 100 and 500 iterations (steps of 100) to identify the optimal numbers oriented towards the coherence value. Our experiments yielded the highest coherence value of around 0.6 for k=12 using 200 iterations, with the scores implying reliable results [78]. Then, we defined the hyper-parameters of the algorithm as $\alpha = k$ and $\beta = 0.01$. Finally, the first author labeled the LDA topics with feasible categories by using an open-card sorting method and added these into the Excel spreadsheet. The first author manually validated all topics, which led to the grouping of five topics (i.e., eight AI incidents) as a new

category we refer to as *Other*—since these were hardly represented. Thus, we relied on a final number of seven topics.

4 RESULTS

In this section, we report our results according to our superordinate topics (i.e., *entry attributes*, *systems*, and *incidents*). Note that the detailed results (i.e., harmonized datasets) can be found in our published replication package.[1]

4.1 Entry Attributes

First, we analyze the entry attributes, via which we aimed to discover the general relevance and trends of incidents connected to safety, security, and both together (cf. Section 5.1). In this context, we focus on the *occurrence years* (cf. Figure 2) and *domains*.

Safety. We identified 132 AI incidents in the context of safety (small sub-bar on the left in Figure 2). These generally represent a growing number of incidents since 2014 (0). Still, from 2015 (5) to 2019 (9), the number of incidents related to safety reported each year were at an almost constant level, with an average of 10 incidents per year. Since 2020 (17), we can observe another step towards more AI incidents being reported every year with 26 incidents in 2021 and 2022, respectively. Interestingly, we found fewer incidents in 2023 (19). We identified a wide range of 14 domains. The most relevant domains are automotive (76), healthcare (14), and transport/logistics (12). Other domains that occur fewer times include manufacturing (6), entertainment (4), general technology (4), or military (3).

Security. Overall, we found 62 incidents related to security (small sub-bar in the middle in Figure 2). Their number has generally been increasing since 2016 (5). In fact, we identified the first reported incidents in 2016, followed by seven in each 2017 and 2019—but with no reported incidents in 2018. Throughout the next two years, we only found two (2020) and four incidents (2021). Surprisingly, the number of security incidents reported has sharply increased for 2022 with 11 incidents and 2023 with 26. The incidents took place in 14 domains, partly in systems that are used in several domains or are domain-independent (8). However, in general, the technology sector (19) dominates the reported incidents. Other relevant domains are consumer goods (8), entertainment (6), finance (6), and research (5). The less mentioned domains include education (2), government (2), and healthcare (2).

Safety and Security. We identified 46 incidents that pose a critical risk to both safety and security at the same time (sub-bar to the right in Figure 2). From 2016 with three incidents (no incidents in 2017), the number of incidents has increased in 2018 (1), 2019 (4), 2020 (7), and 2021 (12). After a slight decrease in 2022 (8), we found 11 incidents in 2023. The incidents cover 14 domains, including technology (7), finance (6), and healthcare (5) as the most important ones. In addition, there are incidents related to automotive (5), business (4), government (4), entertainment (3), and education (3).

4.2 Systems

Second, we analyzed which systems were impacted (i.e., system *name*), which *technologies* they are based on, and which *purpose* they were originally intended to serve. In addition, we present the incident categorization into *working areas* based on our topic modeling (cf. Figure 3).

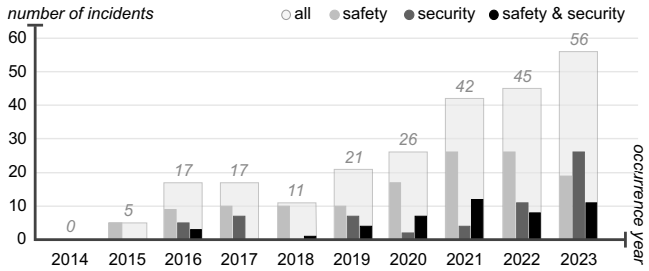

Figure 2: Number of incidents posing critical risks to safety, security, and both safety and security (2014–2023).

Safety. The impacted systems are very diverse. However, the majority of safety incidents occurred in the context of the Tesla driver assistance and autopilot system (50) and Cruise autonomous driving systems (9). Other systems occur significantly less frequently, for example, four cases in the context of Generative Pretrained Transformers (GPT) or three cases in the context of Google navigation applications. We could not specify 29 systems based on the dataset entries. Corresponding to the dominance of automotive systems, the impacted technologies are most often: partly combined computer-vision-related driver assistance systems (50), self-driving systems (29), and robotics (17). Other technologies refer to large-language models and natural-language processing (9) or use case-specific deep-learning methods based on artificial neural networks (6). Consequently, common purposes include automated steering, acceleration, and braking (72) as well as generating text to provide information (9). Interestingly, we also found safety incidents related to the prediction of health states (6).

Topic modeling: Most of the AI systems analyzed (76) are related to automated vehicles. Furthermore, there are several incidents related to the topics of robotics systems (21), prediction systems (17), generative AI (12), and automated air vehicles (4).

Security. The most dominant systems in the context of security are ChatGPT (14) and Apple FaceID (5). Further systems are quite widespread, including chatbots like Microsoft Tay (1) or Google Bard/Gemini (1), diverse face recognition systems like FaceTag (1), or banking systems like the HSBC voice recognition system (1). Six systems are unclear or unknown. Thus, the common technologies refer mainly to large-language models and natural-language processing (21) as well as facial recognition (14). Other technologies are based on generative adversarial networks, for example, in the context of deepfakes (6), as well as use case-specific artificial neural networks (6). We found that the main purposes are oriented towards generating text to provide information (17) as well as strengthening the security of a certain system (14). Other purposes involve defrauding (3), image generation (3), or identification of people (3).

Topic modeling: We classified working areas mainly in the context of generative AI (29), computer vision (18), as well as monitoring (5) and prediction systems (4). In addition, six systems are categorized as "Other," for instance, including financial service systems (2) and development systems (2).

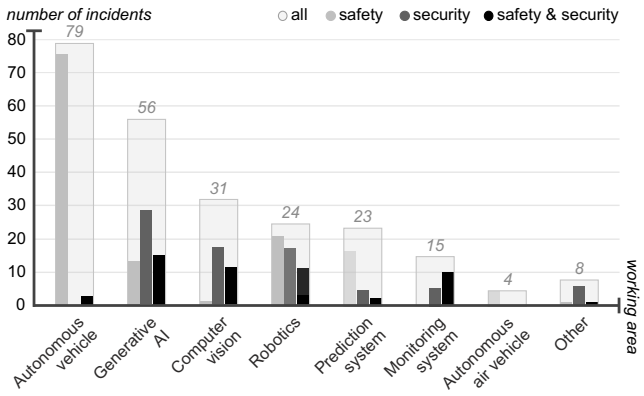

Figure 3: Number of incidents in each working area based on the topic modeling results, classified according to safety, security, and both safety and security.

Safety and Security. The systems that are impacted by an incident connected to safety and security are typically not specified (25) or are highly diverse. For instance, these systems include the Uber driving system (2), ElevenLabs (2), or the iProov face verifier (1). Interestingly, we found that most systems rely on technologies related to facial recognition (13) or generative adversarial networks for deepfakes (12). Other systems relate to location tracking algorithms (2) or decision-support systems (2). Consequently, common purposes include defraud (17), tracking and monitoring of persons, or strengthening system security (2). The less frequently mentioned purposes are the prediction of actions (1) or the increase in operating efficiency in data sharing (1).

Topic modeling: Typical working areas that we extracted are generative AI (15), computer vision (12), and monitoring systems (10). Furthermore, we found systems related to robotics systems (3), automated vehicles (3), and prediction systems (2).

4.3 ⚠ Incidents

Third, we analyzed the incidents and their characteristics. In particular, we focus on the *origin* of the incident, the *risks* that arose, if there were any *serious consequences*, and the actual *harms*. Note that, especially in the context of harms, information is usually missing, as it is often omitted from the sources for legal reasons. Thus, we can assume that the trends in the context of harms are actually more severe than it seems (e.g., economic loss, system updates).

⚠ **Safety.** We found that most incidents related to safety (123) are caused by misbehavior of an AI system. Examples for this are the crash of a Tesla Model S (e.g., `AIAAIC0188`) or faulty predictions of healthcare AI systems (e.g., `AIAAIC0657`). The remaining incidents (9) refer to misused AI issues, for instance, delivery drivers who are forced to take unsafe routes due to an AI-driven navigation algorithm (e.g., `AIAAIC0753`). Besides threatening safety, these incidents also pose risks regarding outcome accuracy and reliability (104), robustness (12), the environment (11), and ethical concerns (10). Interestingly, in 91 cases, the incidents had serious consequences, while only 41 cases did not actually result in critical safety-related harm—or could be prevented in time. The incidents caused several external harms, including 30 cases of fatalities (e.g., `AIAAIC0586`), 26 cases of injuries (e.g., `AIAAIC0918`), and nine cases of substantial economic loss (e.g., `AIAAIC0677`). Moreover, there are internal harms, which are mainly related to 55 cases pf regulatory investigations (e.g., `AIAAIC1176`). Other such harms include system updates (12) or system suspension (6), for example, in the context of a security robot that hit persons (e.g., `AIAAIC062`).

⚠ **Security.** In the context of security-related incidents, we found 31 cases each for misbehavior of the AI system and misused AI issues. For example, there was a misconfiguration of the facial recognition software ClearviewAI (i.e., misbehavior) resulting in authentication issues (`AVID-2023-V007`) or a GPT-based Twitter chatbot that was misused for hijacking purposes (i.e., misused AI issue) via prompt injection attacks (`AIID-352`). The identified incidents pose particular risks regarding accuracy and reliability (29) as well as privacy (28). Moreover, there are also several ethical concerns (9). Overall, 37 incidents are related to the exploitation of certain weaknesses, for example, misusing ChatGPT to perform remote code executions (`AVID-2023-V027`) or backdoor attack on deep-learning models in mobile apps (`AVID-2023-V013`). In 25 cases, the existing weaknesses were not maliciously exploited but detected, for example, a vulnerability in the SenseNets facial recognition system (`AIAAIC0196`) or Apple Face ID failure due to nearly identical persons (`AIAAIC093`). External harms mainly relate to privacy loss (6), data breaches (6), and economic loss (6). Examples for these are the misuse of MathGPT for code executions via prompt injection (`AVID-2023-V016`) or the creation of an AI impersonation to scam people resulting in thousands of USD loss (`AIAAIC1006`). Furthermore, internal harms mainly refer to system updates (10) and system suspension (5), for instance, camera hijacking on a facial recognition system that was updated to solve the weakness (`AVID-2023-V005`). For 38 incidents, harms are not further specified.

⚠ **Safety and Security.** Incidents that pose risks to both safety and security are slightly more related to misbehavior of the AI systems (25) compared to misused AI (21). Such incidents include the facial recognition system Everseen (i.e., AI misbehavior) that comprised several serious bugs (`WitWiAI-03`) or the misuse of AI-generated audio deepfakes (i.e., misused AI) to threaten people and gain access to sensitive data (e.g., `AIAAIC1117`). The identified incidents pose additional risks to accuracy and reliability (14), privacy (13), and ethical concerns (10). In 32 cases, malicious actors exploited weaknesses or caused serious safety-related consequences. For example, biometric cameras were hacked, resulting in serious privacy and safety concerns (e.g., `AIAAIC0507`). We found only 14 incidents with no serious consequences, such as unintended sharing of healthcare patient data (e.g., `AIAAIC0647`). While concrete harms are typically not specified (33), there are several cases regarding external harms including economic loss (6), manipulation (6), and privacy loss (5). For instance, there are various cases in which deepfakes were used to manipulate authentication systems or people and extort money, leading to privacy violations and theft (e.g., `AIAAIC1020`). Also, we found incidents that led to regulatory investigations and litigation (7), system updates (5), and system suspensions (3). For example, a Tesla Model S was remotely controlled by hackers (`AIAAIC067`) and a hotel robot showed serious security vulnerabilities (`AIAAIC0681`).

5 DISCUSSION

In this section, we discuss the results of our analysis to answer our research questions. Precisely, we first analyze our chronological overview of the distribution of AI incidents related to safety, security, and their intersection (RQ_1). Then, we study the most relevant causes and contexts of common AI incidents (RQ_2).

5.1 RQ_1: Topic Relevance

Not surprisingly, our results underpin an increasing number of AI incidents related to safety, security, and their intersection. In turn, this shows the growing relevance, and thus importance, of the topic to the research and development communities. When comparing the years 2020 and 2023, the number of all incidents has actually almost doubled within these four years (cf. Figure 2). Seeing this trend and the continuously growing use of AI systems, we assume that the number of incidents is likely to continue to rise considerably in the coming years. Interestingly, most contemporary incidents relate to safety (132), followed by security (62), and the intersection between both (46).

Regarding safety incidents, we found particularly high and constant numbers of incidents also from 2020 to 2023. This trend is mainly due to incidents related to the automotive domain, for example, in autonomous driving. Precisely, between 2021 and 2023, there were 45 incidents in the automotive domain, which accounts for approximately a third of all incidents (34 %) since 2015. We argue that this dominance has also implications for the distribution of the incidents related to the intersection of safety and security, for example, when components of safety-critical cyber-physical systems (e.g., vehicles) were hacked and resulted in safety violations. So, the intersections of safety and security incidents underpin the complexity and potential multifaceted risks associated with AI.

Interestingly, the number of security incidents has significantly increased only from 2022 to 2023. In particular, we found incidents and challenges related to the spread of one new technology: generative AI—which has had a huge impact on the AI-systems landscape. Specifically, incidents with generative AI from 2023 onward account for 27 % of all security incidents in the last ten years. This finding matches results of recent studies that highlight great security risks of generative AI, especially related to the fulfillment of relevant security goals (e.g., CIA triad) [14, 30]. In fact, 2023 is also the first year with more reported security than safety incidents. Consequently, we can observe a slight shift from a high number of safety incidents to more security incidents and those related to both. Moreover, 2023 is also the first time in five years at which the number of security incidents as well as incidents in the intersection of safety and security together is higher than those incidents that cause safety risks only. This fact emphasizes the emerging role of security in the context of AI systems, for example, regarding their potential to fail [82] or being misused [13]. Furthermore, it also shows the growing dependencies between safety and security, especially in the context of complex safety-critical systems [51]. Note that additional influences may reinforce the trends we observed all the more, such as the increasing number of functions of such AI systems (i.e. features), their configurations, and (cross-)dependencies [50, 52, 53].

> **RQ_1 – Topic Relevance:** *AI incidents related to safety, security, and the intersection between both are an increasingly relevant topic. While safety incidents dominated between 2015 and 2022, 2023 marked a shift towards more security incidents, emphasizing the emerging and critical role of security, in particular in the context of generative AI.*

5.2 RQ_2: Main Incidents and Contexts

We identified that safety and security incidents typically occur in different domains. Although this fact was not surprising, it shows that AI systems and their specifications for these domains pose high risks, either to safety or security. In detail, regarding safety, most reported risks occurred in the context of the automotive or healthcare domains, while security incidents occurred more often in the general technology sector. For incidents referring to both safety and security, we found that the respective systems are either applicable to multiple domains or refer to domains that are typical for either safety or for security. Interestingly, we observe a similar situation regarding the working areas, their technologies, and associated purposes. So, based on the working areas in which incidents happened, we identified patterns regarding the systems that are particularly frequently affected, and thus pose major risks to safety, security, or their intersection. Note that the working areas are significantly related to the domains, but can also deviate from them (e.g., robotics in healthcare vs. robotics in manufacturing). In Figure 4, we illustrate the patterns we identified based on the two dominant working areas for each risk.

Safety. Regarding safety, typically autonomous vehicles (e.g., cars) in the context of their abilities in assisted driving (i.e., autonomous driving, driver assistance systems) lead to issues. So, the prevalence of computer-vision-related driver assistance systems and self-driving systems (i.e., automated steering, acceleration, braking) implies that incidents often arise from complex machine-learning models involved in decision-making processes. Surprisingly, these issues occur due to unwanted or unexpected errors of the systems themselves. These errors usually lead to serious consequences (78 %) like human fatalities or injuries. Accordingly, these are often problems that actually pose significant risks to people's safety. Our findings in this regard are also in line with current research in the automotive sector [40, 72]. The great risk regarding the systems' reliability (79 %) is one of the main reasons for the much-discussed legal basis and the difficulties in the implementation of these systems in actual traffic [5].

The second pattern refers to robotics systems in process automation (e.g., in manufacturing). Similarly to autonomous vehicles, incidents are highly related to errors in AI systems that lead to serious consequences in most cases (78 %). This problem is currently discussed in the research community, leading to novel approaches based on machine-learning systems to prevent such kind of robotics malfunction [81]. However, we argue that machine-learning systems in such contexts pose additional risks that typically refer to both safety and security. Specifically, the data models of monitoring systems (e.g., condition monitoring) or prediction systems (e.g., predictive maintenance) might be compromised by malicious

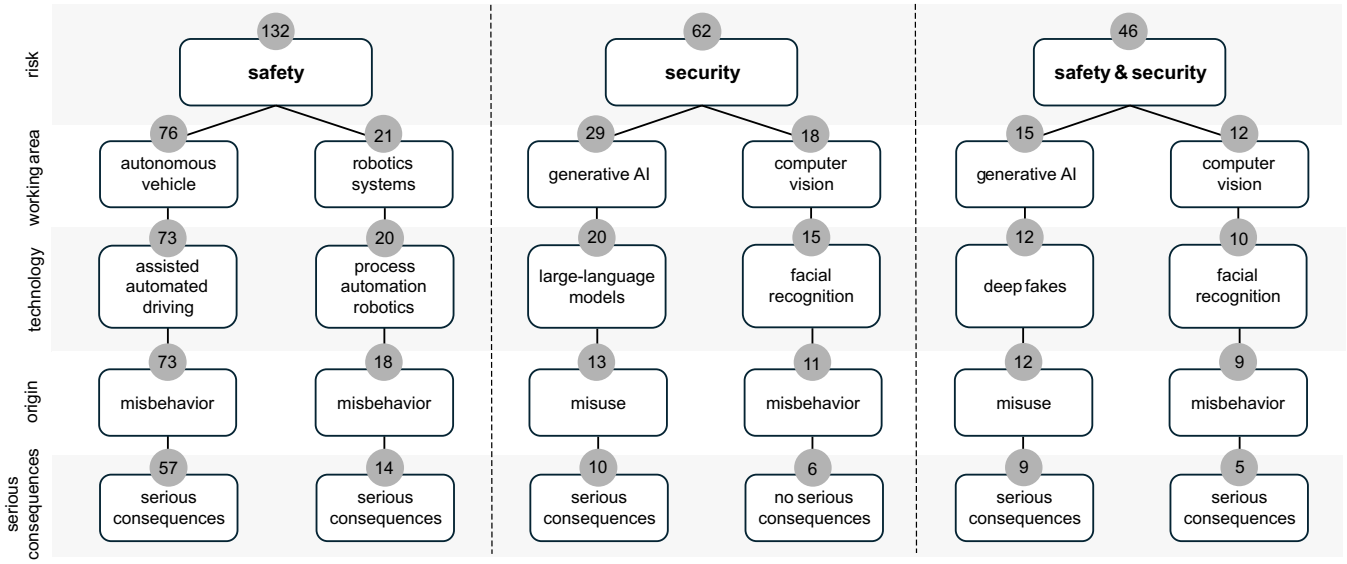

Figure 4: Overview of the most common patterns of AI incidents oriented towards risks, working area, technology, origin, and serious consequences. The numbers indicate the number of AI incidents according to the working areas.

actors [52], potentially leading to economic damage but also human harm. Recent studies have already shown that machine-learning models can involve vulnerabilities that may be exploited by adversarial attacks [43, 64]. These in turn can lead to fatal misclassifications, and thus again AI misbehavior with safety risks.

A working-area-independent perspective of safety incidents (cf. Figure 5) shows that the trend of misbehavior as the cause of issues can generally be related to safety risks. In detail, 93 % of the identified incidents are due to incorrect behavior of AI systems, of which 72 % lead to serious safety violations (e.g., injuries). Surprisingly, the few cases of misuse (7 %) have hardly had serious consequences. These results may indicate that the safeguards in place to prevent misuse or the nature of the misuse itself may be less prone to causing severe harm. So, we argue that there is a critical need for improving the accuracy and reliability of these AI systems, especially those deployed in safety-critical environments (e.g., cyber-physical systems).

> **RQ_2 – Safety:** *Safety incidents, particularly in autonomous vehicles and robotics for process automation, frequently result from system misbehavior and often include serious consequences to both functional and human safety.*

Security. Referring to security, generative AI is the most dominant working area (47 %) leading to incidents. Precisely, these issues are mainly related to large-language models (69 %), for example, large-language model-based systems like ChatGPT. These are usually based on supervised-learning strategies, meaning they do not learn actively, but are based on data that must first be labeled for learning purposes [68]. Large-language models are typically intended to support people (e.g., developers) by providing information to understand topics and solve tasks [36]. Interestingly, in 65 % of the related incidents, their abilities are misused, which typically leads to serious consequences in 77 % of the cases. The possibilities of misuse in the security context are manifold. They range from generating remote code executions (e.g., `AVID-2023-V027`), over hijacking by using prompt injections (e.g., `AIID-352`) to phishing emails in the context of social engineering (e.g., `AIAAIC1211`). In this context, we argue that the misuse mainly jeopardizes the three goals of the CIA triad (i.e., confidentiality, availability, integrity) as well as authorization. Moreover, the capabilities of large-language models are misused both on a technical level and on an ethical level. The use of large-language model-based systems has been subject of much discussion not only in the media but also in research, particularly since the release of ChatGPT in 2022 [54]. However, regulating generative AI based on standards and laws is not trivial and leads to a variety of problems, including limiting the actual potential of the technology [27]. Note that generative AI, in particular ChatGPT, also involves several cases of misbehavior (35 %), especially in the context of privacy violations, which poses additional requirements for the creation and implementation of legal regulations. Consequently, in addition to security, generative AI also entails additional risks in terms of accuracy, reliability, and privacy.

The second pattern we identified is related to the working area computer vision (29 %), in particular facial-recognition systems (83 %). Although facial-recognition systems cover various use cases, they share the common characteristic that they are related to authentication issues (e.g., Apple FaceID). The identified incidents usually refer to misbehavior (73 %), which, however, somewhat more often does not lead to serious consequences (55 %). For instance, bugs in the recognition of faces by smartphones typically do not lead to serious security violations like unlocking a device by a twin (e.g., `AIAAIC099`). Nevertheless, we have found 45 % more serious cases, like successful camera hijacking, leading to unauthorized access and system failure (e.g., `AVID-2023-V005`). In this context,

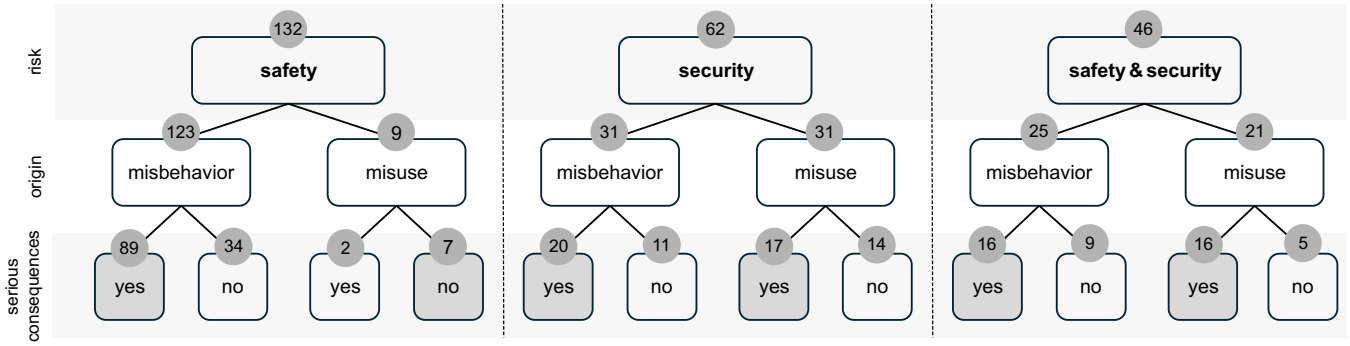

Figure 5: Likelihood of serious consequences (most common trend marked in gray) according to risks and their general origins. The numbers indicate the number of AI incidents.

we argue that there are again major violations of the security goals of the CIA triad (confidentiality, integrity, availability) as well as authorization. Furthermore, the incidents related to facial-recognition systems involve not only security risks, but also risks regarding reliability, but above all privacy, since sensitive personal data is often handled by these systems. Ensuring the correct operation and security of facial-recognition systems is therefore all the more essential, not only to generally prevent risks but also to actively mitigate them in the event of an attack [74].

Overall, AI incidents related to security can be found in the context of both misbehavior (e.g., computer vision) and misuse (e.g., generative AI) due to the systems' technological diversity (50 % in each case). However, as we display in Figure 5, both origins generally lead to serious consequences, and thus serious violations of security and privacy. This finding highlights once again the close connection between security and privacy.

> **RQ_2 – Security:** *Security incidents, especially in generative AI (i.e., large-language models) and computer vision (i.e., facial-recognition systems), involve both misuse and misbehavior, mainly leading to serious consequences and jeopardizing systems' confidentiality, integrity, availability, and authorization.*

Safety and Security. Incidents related to both safety and security violations are quite diverse, indicating a greater independence from domains and working areas. However, we identified two patterns that occur more frequently. Interestingly, both refer to the same working areas like the incidents that solely relate to security. The first pattern is related to generative AI (33 %) for the generation of deep fakes (80 %). These deep fakes typically involve images, audio, and videos. In all cases, the deep fakes are misused for criminal purposes with 75 % resulting in serious consequences. Typically, criminals use deep fakes in the context of defraud and unauthorized access, implying that in particular authentication mechanisms are jeopardized. So, we argue that there are particularly violations of the security goals confidentiality, integrity, authorization, and non-repudiation. Furthermore, due to the misuse of deep fakes, several harms occur for affected people or companies, including serious safety threats, loss of personal rights, or even economic damage (e.g., `AIAAIC1005, AIAAIC1117`). The prevention of a possible system compromise via deep fakes is already being discussed in research, involving several approaches that focus on the effective detection of AI-generated images, audio, or videos [57]. In the context of defraud or even extortion, the regulation of generative AI in particular is a topic that is linked to numerous ethical issues [56].

The second pattern refers to the computer vision working area (26 %) and is quite similar to the second pattern of the incidents that related only to security. Precisely, 83 % of these systems are facial-recognition systems whose misbehavior (90 %) leads to serious consequences (56 %); jeopardizing confidentiality, integrity, availability, and authorization. We emphasize that the patterns differ in particular in the fact that serious consequences occur somewhat more frequently in the context of safety and security. Nevertheless, we argue that the patterns relate to quite similar systems at technological level, for example, in the context of identity verification. The difference here, however, is that these systems also affect people's safety through their misbehavior. For example, the incorrect identification of people during the COVID-19 pandemic led to infected people coming into contact with healthy people (e.g., `AIAAIC0593`). This example highlights the connection between security and safety, namely the need for correct and reliable security mechanisms (i.e., authentication) to ensure human safety. In Figure 5, we display the general considerations of all incident consequences related to the intersection of safety and security, which shows that these are mostly serious (70 %). This finding is evident for both misbehavior (54 %) and misuse (46 %).

> **RQ_2 – Safety and Security:** *The intersection of safety and security incidents (i.e., deep fakes, facial recognition) often results in serious consequences, emphasizing the close connection between security and safety, with implications for system reliability, confidentiality, integrity, availability, authorization, as well as human safety.*

When comparing incidents related to security, safety, and their intersection, several contexts emerge. While incidents that only occur in the context of safety mostly relate to functional safety (i.e., damage to environments, systems, users), they can also impair human safety (i.e., human injuries). Moreover, they are typically caused by both software and hardware failure (i.e., misbehavior).

In the case of additional risks regarding security, these are usually exclusively related to human safety and software failure. Furthermore, we argue that safety incidents mostly relate to the automation of hardware (e.g., steering, braking), while security adds a strong context to or even replaces this focus with authentication. Interestingly, we found only very few cases where both safety and security are jeopardized in the context of automated (safety-critical) systems, such as a car which was remotely controlled by hackers (`AIAAIC067`). Possible reasons given in the datasets are diverse (e.g., non-disclosure of information to avoid economic damage), which is why we strongly recommend further research in the intersection of safety-critical AI systems and security.

6 THREATS TO VALIDITY

There are several threats to the internal validity (i.e., correct attribution of data to variables), external validity (i.e., generalizability of findings), and construct validity (i.e., accuracy in representing theoretical concepts with data) of our study which we discuss in the following.

Internal Validity. There may be threats in the contexts of misinterpreting the AI incidents as well as the reports they reference. We encountered AI incidents with different levels of detail and terminologies due to the diversity of authors as well as responsible persons of the datasets (e.g., different nationalities). This fact led to various interpretations on our side. In addition, we also made adjustments within the existing datasets to use standardized terms and enable a better overall understanding. We addressed these threats by applying strict selection criteria to ensure an appropriate quality as well as a clear connection to our research goal and its associated main field (i.e., security, safety). Moreover, we validated our classifications and performed LDA, which provided suitable topics, particularly for the domains, technologies, and purposes. This, in turn, also mitigated the threat of misclassifications.

External Validity. The external validity is threatened by the fact that we may have missed relevant incidents, for example, due to incidents that we deleted because they seem identical—but that actually comprise (slightly) different characteristics (e.g., incidents related to automated driving). In addition, we only relied on AI incident datasets that mainly refer to media reports, usually without scientific foundation. The datasets may not represent all relevant incidents that actually occurred, since they are oriented towards media interest (cf. construct validity). A comparison with other datasets, for example, the National Vulnerability Database in the context of security vulnerabilities, could have provided additional credibility here. However, we argue that we based our study on the four largest AI incident datasets, whose high data quality is also validated by recent scientific literature [48, 67, 82]. Furthermore, we based our study on a sufficient high number of incidents (240), which allows the extraction of partly expected trends (e.g., risks of generative AI). Nevertheless, we are aware that an even higher number of incidents would have been contributed towards providing even more generalizable conclusions.

Construct Validity. Although there are several incidents that are reported in scientific studies, for example, regarding black-box backdoor attacks on deep-learning models [44], most incidents we studied are based on articles from reputable newspapers. These typically rely on the opinions of their authors who do not write through the lens of safety and security, implying potentially biased data. The same applies to the responsible persons of the datasets, who, however, probably have more experience in the context of AI, safety, and security. Even though we are generally unable to verify who exactly maintains the incidents, we trust the quality controls of the owners, which are stated on their websites and in associated scientific literature [55, 67, 82]. Although considering four datasets resulted in broader insights, these rely on different data structuring (i.e., classifications). Moreover, the descriptions of the incidents and the reports themselves are completely unstructured. However, the lack of structure also offered advantages, such as the possibility to apply LDA as well as to obtain broader perspectives and opinions, which are roughly comparable to conducting interviews with people who are interested in AI systems [77]. Furthermore, we read all AI incidents and associated reports and manually validated all selected incidents to ensure that the classifications are correct and understandable. So, we argue that the data has been checked several times with regard to its structure and overall quality.

7 RELATED WORK

Extensive research has been conducted at the intersection of safety, security, and AI systems. It can be roughly classified into work using AI methods for safety or security purposes as well as ensuring safety and security for AI systems to overcome typical risks. The former comprises diverse publications, for example, artificial neural networks to analyze general security capabilities [66] or to detect security risk factors in cloud computing [80], or using AI to identify safety hazards in the construction industry [33]. The latter refers more to potential risks, challenges, and incidents related to AI, for example, security attack vectors of AI systems [63] or critical patient safety in AI-driven medical environments [12].

There is only little research that is based on systematic analyses of AI incidents datasets, which would be similar to our work regarding methodology and data sources. Currently, most studies rely either on the AIAAIC repository or the AIID. Burema et al. [9] conducted an analysis of 125 AI incidents of the AIAAIC repository. They focused on AI ethics from a domain perspective (e.g., automotive, healthcare). Das et al. [15] examined 321 AI privacy incidents based on the AIAAIC repository. Their data sample partly overlaps with our data on security, since privacy concerns may relate to security issues (e.g., training data protection). Moreover, Golpayegani et al. [24] analyzed 52 entries of the AIAAIC repository that are related to healthcare AI systems. They created a catalog of AI risks, sources, consequences, and their impact. Stanley and Dorton [77] investigated 30 incidents of the AIID related to loss of trust. Interestingly, the authors mentioned security concerns as one relevant factor in losing trust in AI systems. Lastly, Nasim et al. [61] analyzed 155 incidents of the AIID, focusing on ethical concerns.

In contrast to the related work, our study focuses on safety, security, and their connections in the context of AI incidents. In addition, we combined and harmonized the four currently largest datasets on AI incidents oriented towards the AIAAIC repository categories, giving us access to a greater body of knowledge. Thus, we argue that we have conducted the most comprehensive work in the field of AI incidents dataset analysis so far, which is of particular

value to the research community. Overall, we contribute novel insights from a practice-oriented perspective that have not been reported in the related work.

8 CONCLUSION

In this paper, we presented a case survey based on reports from four AI incidents datasets, namely the AIAAIC repository, the AIID, the AVID, and WitWiAI. We focused on AI incidents in the context of safety, security, and their intersection that occurred during the past decade. So, we provide an overview of common AI incidents in these fields as well as an understanding of the connections between AI systems, safety, and security. The main findings of our study are:

- AI incidents related to security, safety, and their intersection are an increasingly relevant topic that is currently slightly shifting from more safety issues to more security issues, including security cases with greater safety implications.
- Safety incidents (i.e., functional and human safety) mainly occurred in the context of autonomous vehicles and robotics for process automation, which frequently result from system misbehavior and caused serious consequences (e.g., injuries).
- Security incidents are particularly related to generative AI misuse (i.e., large-language models) and computer vision (i.e., facial-recognition systems) misbehavior.
- Serious consequences related to security violations mostly relate to a system's confidentiality, integrity, availability, and authorization, typically in authentication contexts.
- Incidents in the intersection of safety and security typically stem from generative AI misuses (i.e., deep fakes) and computer vision (i.e. facial recognition), mainly leading to serious consequences in the context of human safety.

Based on our results, several implications for future research emerge. In the context of safety and security, these include further investigations of safety-critical AI systems and security, for instance, analyzing which additional attack vectors are created by the AI capabilities (e.g., system evolution) and how these influence the risks associated with misbehavior or misuse. In addition, we recommend to extend our results by comparing security-related incidents with entries of vulnerability datasets (e.g. National Vulnerability Database). As a result, more in-depth information and associated trends can be identified, in particular regarding the affected systems and their configurations.

REFERENCES

[1] A. Agrawal, W. Fu, and T. Menzies. 2018. What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology* 98 (2018), 74–88.

[2] I. Al Ridhawi, S. Otoum, M. Aloqaily, and A. Boukerche. 2020. Generalizing AI: Challenges and opportunities for plug and play AI solutions. *IEEE Network* 35, 1 (2020), 372–379.

[3] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[4] L. Banh and G. Strobel. 2023. Generative artificial intelligence. *Electronic Markets* 33, 1 (2023), 1–17.

[5] I. Begishev, D. Bersei, L. Sherbakova, R. Zhirov, and O. Kolesnikova. 2022. Problems of legal regulation of unmanned vehicles. *Transportation Research Procedia* 63 (2022), 1321–1327.

[6] R. Bell. 2006. Introduction to IEC 61508. In *International Conference Proceeding Series*, Vol. 162. ACM.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 1 (2003), 993–1022.

[8] M. F. Bosu and S. G. MacDonell. 2013. Data quality in empirical software engineering: a targeted review. In *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. ACM, 171–176.

[9] D. Burema, N. Debowski-Weimann, A. von Janowski, J. Grabowski, M. Maftei, M. Jacobs, P. Van Der Smagt, and D. Benbouzid. 2023. A sector-based approach to AI ethics: Understanding ethical issues of AI-related incidents within their sectoral context. In *Conference on AI, Ethics, and Society (AIES)*. ACM, 705–714.

[10] Á. A. Cabrera, A. J. Druck, J. I. Hong, and A. Perer. 2021. Discovering and validating AI errors with crowdsourced failure reports. *ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.

[11] J. C. Campbell, A. Hindle, and E. Stroulia. 2015. Latent Dirichlet allocation: Extracting topics from software engineering data. In *The Art and Science of Analyzing Software Data*. Elsevier, 139–159.

[12] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety* 28, 3 (2019), 231–237.

[13] I. Chiscop, F. Soro, and P. Smith. 2022. AI-based detection of DNS misuse for network security. In *International Workshop on Native Network Intelligence (NativeNI)*. ACM, 27–32.

[14] Md M. Chowdhury, N. Rifat, M. Ahsan, S. Latif, R. Gomes, and Md S. Rahman. 2023. ChatGPT: A threat against the CIA triad of cyber security. In *International Conference on Electro Information Technology (eIT)*. IEEE, 1–6.

[15] S. Das, H.-P. Lee, and J. Forlizzi. 2023. Privacy in the age of AI. *Communications of the ACM* 66, 11 (2023), 29–31.

[16] K. Denecke, R. May, and O. Rivera-Romero. 2024. Transformer models in healthcare: A survey and thematic analysis of potentials, shortcomings and risks. *Journal of Medical Systems* 48 (2024), 1–11. Issue 23.

[17] A. Dogan and D. Birant. 2021. Machine learning and data mining in manufacturing. *Expert Systems with Application* 166 (2021), 1–45.

[18] M. Ebnauf, W. Abdelmoez, H. H. Ammar, A. Hassan, and M. Abdelhamid. 2019. State-driven architecture design for safety-critical software product lines. In *ICOM General Conference (ICOM)*. IEEE.

[19] T. E. Fægri and S. Hallsteinsen. 2006. A software product line reference architecture for security. In *Software Product Lines*. Springer, 275–326.

[20] G. Falco, B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, et al. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence* 3, 7 (2021), 566–571.

[21] M. Feffer, N. Martelaro, and H. Heidari. 2023. The AI incident database as an educational tool to raise awareness of AI harms: A classroom exploration of efficacy, limitations, & future improvements. In *Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. ACM, 1–11.

[22] M. I. Ganesh and E. Moss. 2022. Resistance and refusal to algorithmic harms: Varieties of 'knowledge projects'. *Media International Australia* 183, 1 (2022), 90–106.

[23] S. S. Gill, M. Xu, C. Ottaviani, P. Patros, R. Bahsoon, A. Shaghaghi, M. Golec, V. Stankovski, H. Wu, A. Abraham, et al. 2022. AI for next generation computing: Emerging trends and future directions. *Internet of Things* 19 (2022), 100514.

[24] D. Golpayegani, J.. Hovsha, L. W. S. Rossmaier, R. Saniei, and J. Mišić. 2022. Towards a taxonomy of AI risks in the health domain. In *International Conference on Transdisciplinary AI (TransAI)*. IEEE, 1–8.

[25] D. Golpayegani, H. J. Pandit, and D. Lewis. 2022. Airo: An ontology for representing ai risks based on the proposed EU AI act and ISO risk management standards. In *International Conference on Semantic Systems (SEMANTICS)*, Vol. 55. IOS Press, 51.

[26] I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning*. The MIT Press.

[27] P. Hacker, A. Engel, and M. Mauer. 2023. Regulating ChatGPT and other large generative AI models. In *Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 1112–1123.

[28] J. Hatcliff, A. Wassyng, T. Kelly, C. Comar, and P. Jones. 2014. Certifiably safe software-dependent systems: Challenges and directions. *International Conference on Software Engineering – Future of Software Engineering (FOSE)* (2014).

[29] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–16.

[30] D. Humphreys, A. Koay, D. Desmond, and E. Mealy. 2024. AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business. *AI and Ethics* (2024), 1–14.

[31] IEC 61508 2010. *Functional Safety*. Standard. IEC.

[32] ISO/IEC 27000 2018. *Information Technology – Security Techniques – Information Security Management Systems*. Standard. ISO.

[33] H. Jallow, S. Renukappa, S. Suresh, and F. Rahimian. 2023. Artificial intelligence and the UK construction industry–empirical study. *Engineering Management Journal* 35, 4 (2023), 420–433.

[34] M. Janmaijaya, A. K. Shukla, P. K. Muhuri, and A. Abraham. 2021. Industry 4.0: Latent Dirichlet Allocation and clustering based theme identification of bibliography. *Engineering Applications of Artificial Intelligence* 103 (2021), 104280.

[35] P. P. Kalita, M. P. Sarma, and A. P. Saikia. 2023. Artificial intelligence and robots in individuals' lives: How to align technological possibilities and ehtical issues. *Ethical Issues in AI for Bioinformatics and Chemoinformatics* (2023), 119.

[36] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, and E. Hüllermeier. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274.

[37] A. Kenner, R. May, J. Krüger, G. Saake, and T. Leich. 2021. Safety, security, and configurable software systems: A systematic mapping study. In *Systems and Software Product Line Conference (SPLC)*. 148–159.

[38] A. A. Khan, A. A. Laghari, and S. A. Awan. 2021. Machine learning in computer vision: a review. *Endorsed Transactions on Scalable Information Systems* 8, 32 (2021), 1–11.

[39] H. Kim, J. Ben-Othman, L. Mokdad, J. Son, and C. Li. 2020. Research challenges and security threats to AI-driven 5G virtual emotion applications using autonomous vehicles, drones, and smart devices. *IEEE Network* 34, 6 (2020), 288–294.

[40] P. Koopman and M. Wagner. 2017. Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine* 9, 1 (2017), 90–96.

[41] S. Lefcourt and G. Falco. 2023. AI forensics. In *International Conference on Assured Autonomy (ICAA)*. IEEE, 106–114.

[42] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, et al. 2011. Towards fully autonomous driving: Systems and algorithms. In *Intelligent Vehicles Symposium (IV)*. IEEE, 163–168.

[43] Y. Li, M. Cheng, C.-J. Hsieh, and T. C. M. Lee. 2022. A review of adversarial attack and defense for classification methods. *The American Statistician* 76, 4 (2022), 329–345.

[44] Y. Li, J. Hua, H. Wang, C. Chen, and Y. Liu. 2021. Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection. In *International Conference on Software Engineering (ICSE)*. IEEE, 263–274.

[45] Y. Li and Q. Liu. 2021. A comprehensive review study of cyber-attacks and cyber security: Emerging trends and recent developments. *Energy Reports* 7 (2021), 8176–8186.

[46] Y. Lu. 2019. Artificial intelligence: A survey on evolution, models, applications and future trends. *Journal of Management Analytics* 6, 1 (2019), 1–29.

[47] B. Lundgren and N. Möller. 2019. Defining information security. *Science and Engineering Ethics* 25, 2 (2019), 419–441.

[48] G. Lupo. 2023. Risky artificial intelligence: The role of incidents in the path to AI regulation. *Law, Technology and Humans Journal* 5, 1 (2023), 133–152.

[49] R. May, A. J. Alex, R. Suresh, and T. Leich. 2024. Product-line engineering for smart manufacturing: A systematic mapping study on security concepts. In *International Conference on Software Technologies (ICSOFT)*. SciTePress, 1–8.

[50] R. May, C. Biermann, X. M. Zerweck, K. Ludwig, J. Krüger, and T. Leich. 2024. Vulnerably (mis)configured? Exploring 10 years of developers' Q&As on Stack Overflow. In *International Working Conference on Variability Modelling of Software-Intensive Systems (VaMoS)*. ACM, 112–122.

[51] R. May, J. Gautam, C. Sharma, C. Biermann, and T. Leich. 2023. A systematic mapping study on security in configurable safety-critical systems based on product-line concepts. In *International Conference on Software Technologies (ICSOFT)*. SciTePress, 217–224.

[52] R. May., T. Niemand., P. Scholz., and T. Leich. 2023. Design patterns for monitoring and prediction machine learning systems: Systematic literature review and cluster analysis. In *International Conference on Software Technologies (ICSOFT)*. SciTePress, 209–216.

[53] R. May and X. M. Zerweck. 2024. Towards vulnerabilities caused by application configuring: A meta analysis of the National Vulnerability Database. In *Scientific Reports*. 328–332.

[54] Q. P. McGrath. 2024. Unveiling the ethical positions of conversational AIs: a study on OpenAI's ChatGPT and Google's Bard. *AI and Ethics* (2024), 1–16.

[55] S. McGregor. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Conference on Artificial Intelligence (AAAI)*, Vol. 35. ACM, 15458–15463.

[56] E. Meskys, J. Kalpokiene, P. Jurcys, and A. Liaudanskas. 2020. Regulating deep fakes: Legal and ethical considerations. *Journal of Intellectual Property Law & Practice* 15, 1 (2020), 24–31.

[57] Y. Mirsky and W. Lee. 2021. The creation and detection of deepfakes: A survey. *Computing Surveys* 54, 1 (2021), 1–41.

[58] V. Myllärniemi, M. Raatikainen, and T. Männistö. 2015. Representing and configuring security variability in software product lines. In *International Conference Series on the Quality of Software Architectures (QoSA)*. ACM, 1–10.

[59] M. Nadimpalli. 2017. Artificial intelligence risks and benefits. *International Journal of Innovative Research in Science, Engineering and Technology* 6, 6 (2017).

[60] A. Nardi and A. Armato. 2017. Functional safety methodologies for automotive applications. IEEE, 970–975.

[61] S. F. Nasim, M. R. Ali, and U. Kulsoom. 2022. Artificial intelligence incidents & ethics a narrative review. *International Journal of Technology, Innovation and Management* 2, 2 (2022), 52–64.

[62] NIST SP 800-30r1 2012. *Guide for Conducting Risk Assessments*. Standard. National Institute of Standards and Technology.

[63] O. A. Osoba and W. Welser. 2017. The risks of artificial intelligence to security and the future of work. *Perspective* (2017), 1–23.

[64] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas. 2019. A taxonomy and survey of attacks against machine learning. *Computer Science Review* 34 (2019), 100199.

[65] V. Raina, S. Krishnamurthy, V. Raina, and S. Krishnamurthy. 2022. Natural language processing. *Building an effective data science practice: A framework to bootstrap and manage a successful data science practice* (2022), 63–73.

[66] B. S. Rawat, V. Gangodkar, D.and Talukdar, K. Saxena, C. Kaur, and S. P. Singh. 2022. The empirical analysis of artificial intelligence approaches for enhancing the cyber security for better quality. In *International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 247–250.

[67] R. Rodrigues, A. Resseguier, and N. Santiago. 2023. When artificial intelligence fails: The emerging role of incident databases. *Public Governance Administration and Finances Law Review* 8, 2 (2023), 17–28.

[68] K. I. Roumeliotis and N. D. Tselikas. 2023. ChatGPT and Open-AI models: A preliminary review. *Future Internet* 15, 6 (2023), 192.

[69] S. Samonas and D. Coss. 2014. The CIA strikes back: Redefining confidentiality, integrity and availability in security. *Journal of Information System Security* 10, 3 (2014).

[70] G. Schuh and P. Scholz. 2019. Development of a framework for the systematic identification of AI application patterns in the manufacturing industry. In *International Conference on Management of Engineering and Technology*. IEEE, 1–8.

[71] G. Schuh, P. Scholz, T. Leich, and R. May. 2020. Identifying and analyzing data model requirements and technology potentials of machine learning systems in the manufacturing industry of the future. In *International Scientific Conference on Information Technology and Management Science (ITMS)*. IEEE, 1–10.

[72] S. Shafaei, S. Kugele, M. H. Osman, and A. Knoll. 2018. Uncertainty in machine learning: A safety perspective on autonomous driving. In *International Conference on Computer Safety, Reliability and Security (SAFECOMP)*. Springer, 458–464.

[73] T. Shaffer Shane. 2023. AI incidents and 'networked trouble': The case for a research agenda. *Big Data & Society* 10, 2 (2023), 1–6.

[74] S. Shankar, J. Madarkar, and P. Sharma. 2020. Securing face recognition system using blockchain technology. In *Machine Learning, Image Processing, Network Security and Data Science*. Springer, 449–460.

[75] C. Sharma, S. Sharma, and Sakshi. 2022. Latent Dirichlet Allocation (LDA) based information modelling on blockchain technology: A review of trends and research patterns used in integration. *Multimedia Tools and Applications* 81, 25 (2022), 36805–36831.

[76] M. Sharma, J. Anuradha, H. K. Manne, and G. S. C. Kashyap. 2017. Facial detection using deep learning. In *IOP Conference Series: Materials Science and Engineering*, Vol. 263. IOP Publishing, 042092.

[77] J. C. Stanley and S. L. Dorton. 2023. Exploring trust with the AI incident database. In *HFES Annual Meeting*, Vol. 67. SAGE Publications, 489–494.

[78] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. 2012. Exploring topic coherence over many models and many topics. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 952–961.

[79] A. Strowel. 2023. ChatGPT and generative AI tools: Theft of intellectual labor? *International Review of Intellectual Property and Competition Law* 54, 4 (2023), 491–494.

[80] D. A. G. Tadeo, S. F. John, A. Bhaumik, R. Neware, N. Yamsani, and D. Kapila. 2021. Empirical analysis of security enabled cloud computing strategy using artificial intelligence. In *International Conference on Computational Science (ICCS)*. IEEE, 83–85.

[81] A. Terra, H. Riaz, K. Raizer, A. Hata, and R. Inam. 2020. Safety vs. efficiency: AI-based risk mitigation in collaborative robotics. In *International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, 151–160.

[82] V. Turri and R. Dzombak. 2023. Why we need to know more: Exploring the state of AI incident documentation practices. In *Conference on AI, Ethics, and Society (AIES)*. ACM, 576–583.

[83] T. Wuest, D. Weimer, C. Irgens, and K. D. Thoben. 2016. Machine learning in manufacturing: Advantages, challenges, and applications. *International Journal of Production Research and Manufacturing Research* 4, 1 (2016), 23–45.

[84] P. Zellmer, L. Holsten, R. May, and T. Leich. 2024. A practitioners perspective on addressing cyber security and variability challenges in modern automotive systems. In *International Working Conference on Variability Modelling of Software-Intensive Systems (VaMoS)*. ACM, 129–133.

[85] L. Zhang, J.-H. Tian, J. Jiang, Y.-J. Liu, M.-Y. Pu, and T. Yue. 2018. Empirical research in software engineering—a literature survey. *Journal of Computer Science and Technology* 33 (2018), 876–899.

[86] M. Zwilling, G. Klien, D. Lesjak, Ł. Wiechetek, F. Cetin, and H. N. Basim. 2022. Cyber security awareness, knowledge and behavior: A comparative study. *Journal of Computer Information Systems* 62, 1 (2022), 82–97.