AI Security and Privacy – A State-of-the-Art Review

BENEDIKT REIN, LEILA ZAFARMAND, and THEODOR RELLIN, Humboldt University Berlin School of Business and Economics, Germany

***Abstract* - This systematic mapping study explores recent developments in the area of AI security and privacy and corresponding frameworks. We present some well-established frameworks and evaluate a collection of 75 scientific papers selected by a systematic approach. The frameworks are designed to help practitioners, managers, auditors and policymakers make informed decisions on AI security and privacy. The *EU AI Act* is used as a normative starting point for what society expects from AI systems. We are able to derive lifecycle- and value-based mapping categories from the frameworks. We map the relevant literature to the defined categories and try to explore underlying structures such as contribution type, year of publication and more. Recent topics are federated learning, privacy-preserving training and adversarial attacks. We summarize all lifecycle steps into *AI system architecture, data management, deploy - predict - monitor, model training*, and finally *retirement*. Not as prominent as expected is research into bias management and alignment. From the lifecycle side, we did not find literature discussing retirement in detail. Overall we learned, that AI security and privacy is a very new research field. Governments worldwide are working on effective measures to ensure AI security and privacy to create a safe environment and minimize potential risks. With different societal values, there is no one-size-fits-all solution for AI policy and no framework claims to be a complete blueprint on how to manage AI security and privacy.**

Additional Key Words and Phrases: AI, Security, Privacy, Frameworks, EU AI Act, Policy, Safety, Alignment, AI RMF, Responsible AI, Robustness, Artificial Intelligence, Federated Learning, AI Risk Management Framework, Lifecycle

1 INTRODUCTION

Artificial Intelligence is a fast-growing topic in many research fields. Given the growing application of Artificial Intelligence in different domains such as health care [26] and banking [21], the need for secure and private AI has become more obvious. While more use cases for using AI are being published every day, security concerns are also rising with more powerful AI solutions.

The importance of private data and reliable results has prompted the Standards Developing Organizations to focus on creating standards and frameworks for data collection and processing and model enhancement in AI to address this concern. Many of the proposed frameworks or methods by researchers are built on such standards and guidelines. The European Union has published the *EU AI Act* and a US governmental entity, the National Institute of Standards and Technology (NIST), has published the *AI Risk Management Framework* (AI RMF) which will be discussed more deeply in the following parts of this paper. Data privacy, increasing amounts of data, explainability, malicious use and adversary attacks all call for exploring and defining security risks with the use of AI. Sensitive information makes up a significant portion of the data used to train AI systems, making any vulnerability a major threat to privacy and security. Nonetheless, the growing demand for high-quality predictive or generative software from governments and companies necessitates the use of AI systems. Various studies have outlined the risks posed by big data on different platforms and proposed countermeasures to mitigate these dangers.

Some research addresses the different types of attacks to manipulate the data or the result, while other research aims to propose new frameworks and best practices for secure AI systems.

Authors' address: Benedikt Rein, reinbene@hu-berlin.de; Leila Zafarmand, leila.zafarmand@student.hu-berlin.de; Theodor Rellin, rellinth@hu-berlin.deHumboldt University Berlin, School of Business and Economics, Berlin, Berlin, Germany.

We selected to conduct a systematic mapping since it is better fitting to our approach than a systematic literature review. Mapping studies are especially useful for structuring and organizing emerging areas of research [22].

In our systematic mapping study, our emphasis is directed toward best practices for ensuring security and privacy throughout the AI-based software products and the dimensions within the AI lifecycle with their associated security best practices as our research goal. Our concrete research question is the following: „What are the key frameworks used to manage the technical and societal implications of Artificial Intelligence technologies and are they in line with recent developments in AI security and privacy research?“ We identified, classified, and systematically reviewed the existing literature body on AI security frameworks with the purpose of better understanding and suggesting measurements and frameworks for a secure AI.
We have conducted a systematic mapping study by reviewing studies and selecting 75 of them with an iterative process and by applying inclusion and exclusion criteria. The results of our literature study are mainly focused on approaches, methods, and frameworks for a secure AI considering the AI lifecycle. Furthermore, the results of our study revealed that the area of AI security has attracted growing attention and interest.

This paper is organized as follows: In section 2 we review background and related studies to position this work. In chapter 3 we present our research scope, methodology and classification scheme for the systematic mapping. Section 4 provides the result of our mapping study and chapter 5 describes a deeper discussion of the proposed frameworks. Section 6 concludes the paper and summarizes our key findings.

2 RELATED WORK

There are many different topics surfacing in the field of AI security, safety and privacy. Be it technical solutions, socio-technical discussions or frameworks that intend to cover AI security as completely as possible. Other research proposes frameworks for subtopics, while some discuss specific topics in more depth. Because AI systems are always a trade-off between security, capability, speed, alignment, and many other aspects, it is difficult to define clear, non-biased and objective goals and performance values. Many of those aspects are discussed and defined in research-, company-, institute-, or country-specific contexts. To start from a normative position that is easy to argue for, because the European Union is known for its comparably restrictive cybersecurity and privacy laws, the *EU AI Act* is used as an initial normative baseline, to understand what is expected of AI systems. It is particularly relevant for Germany because the policy will soon be incorporated into German law as well.

After an introduction to the *EU AI Act*, we will give an overview of some well-established AI security frameworks. Entities providing those frameworks are a mixture of companies, government entities, standards organizations and research institutes.

As a first step, we derive a high-level grouping, to create a better understanding of the topic. General topics are the overall technical security and robustness of an AI system with a special focus on many types of adversarial attacks and how to implicitly and explicitly defend against those. Other topics would be the alignment of those systems, bias, and privacy aspects. Two of the most widely used analytical perspectives are a value-based approach and a lifecycle-based approach, those takes will also be used as an initial blueprint for this mapping study, discussed in the following chapter. The topics can be discussed as a full AI security lifecycle framework, from inception to retirement stage, but also more from an MLOps perspective. Which also cares about similar aspects, but takes a more technical perspective. In the systematically explored literature, one or more of the framework topics are discussed in depth. To get a better

understanding, the frameworks are a very insightful starting point.

There are different intentions for making those full frameworks. With frameworks being developed by the *Organisation for Economic Co-operation and Development* (OECD) and the U.S. entity *National Institute of Standards and Technology*, initial goals seem to be to define universal standards on how to define, measure and communicate about AI security but also further the debate about AI security policies and compliance mechanisms. Just recently, Google adapted the NIST framework, publishing their *Secure AI Framework* (SAIF). Defining different perspectives from which to apply lifecycle-, value-, and other approaches. Companies, on the other hand, care more from the perspective of how to make sure, that models are in line with policy. This matters for the inception, design, training, and inference stage. At every point, compliance has to be assured with auditing, privacy, security, and alignment policies. Additionally, companies care about their public perception, making it inevitable to make sure, that models are aligned with societal values and fit for purpose.

Thus, for related work, we take the *EU AI Act* as a normative starting point, for what goals and limitations are or should be set for AI systems. Secondly, we will describe AI security frameworks, that claim a somewhat holistic approach. In the following chapter, we will discuss, how different approaches and perspectives can be structured sensibly to be used for a systematic mapping study that fits our research interests.

2.1 EU AI Act

With this being a computer science paper, an evaluation of the *EU AI Act* from a regulatory law perspective is out of its scope. Thus, we will use secondary literature for a short summary and evaluation of the policy. Veale and Borgesius (2021) did an evaluation of an initial version of the *EU AI Act*, while also summarizing the most crucial points and pointing out changes that have been made compared to previous working papers and leaked versions of the document. It is important to note, that the policy itself is far from being a complete AI policy. Especially the *General Data Protection Regulation* (GDPR) should be mentioned, which is one of the most restrictive data privacy policies. Additional policies mentioned, that work in conjunction with the *EU AI Act* are: the draft *Digital Services Act*, the draft *Digital Markets Act*, the draft *Machinery Regulation*, and the draft *Data Governance Act* [28]. The key feature of the new policy is defining multiple threat levels of AI systems and formulating guidelines to distinguish them and how each category's specific risks and chances should be managed. It should be mentioned, that policy decisions are guided by culture-specific values and the presented perspective is a European one [11].

One of the most discussed points of the EU AI Act is the prohibition of **unacceptable-risk** AI services [28]. Those would be systems that manipulate people unconsciously, social scoring systems, and real-time biometric tracking systems for use by law enforcement. The mentioned paper criticizes many loopholes from a law perspective and states the possibility of intentional loopholes and strong influence by industry. With prohibitions being taken especially seriously, for example in the U.S., this might be a first sign of some incompatibility of values in the future.

The second category, **high-risk systems**, would be AI systems used in infrastructure and physical products or systems, that are used in high-impact areas such as education, law enforcement, biometric identification, employment, or worker management or administration [2]. For those products, the provider of the AI systems is responsible for making sure it is operating in a safe and law-abiding manner. To comply, actors have to implement a *risk management system*, meet *data quality criteria*, manage discrimination or bias, enter information about their system in a central

database managed by the European Commission, and keep in line with expectations concerning *accuracy, robustness, and cybersecurity* [28]. Those requirements make the need for AI lifecycle frameworks apparent, to be able to develop, operate, and audit those systems in a safe and compliant manner. Finally, the policy requires logging of model inputs, outputs and options for human oversight. An interesting observation made by Veale and Borgesius (2021) is, that there are exemptions for using ethnicity data for high-risk applications, which would usually be strictly forbidden in the EU. The reasoning is, that you can just check and evaluate for e.g. racial bias when that information is present in the dataset. This underlines the very suiting category definition in some frameworks as *managed bias*, only when bias is measured and managed, it can be proven to be nonexistent in a model or dataset.

An additional category with specific requirements is **Generative AI systems**, which humans are interacting with or consume content from. AI-based chat systems need to be disclosed, and AI systems need to inform when using any biometric information. Content that has been created by AI needs to be disclosed [2]. Many of the stated requirements can be found in different ways in AI security and privacy frameworks. Those frameworks have been created by researchers, companies, or government and standards organizations. In the following, we will give a brief overview of the most widely used frameworks, how they came to be, and what the important aspects of each framework are. Many frameworks build on top of each other, to prevent repetitions we will mention this for each case and focus on novel aspects of each framework.

2.2 National Institute of Science and Technology

We will start off with frameworks defined by standards organizations and government entities. Currently, one of the most discussed and used frameworks is the second draft of the *AI Risk Management Framework*, published by the *National Institute of Standards and Technology*, which is an entity of the *U.S. Department of Commerce* [18].
The NIST framework takes two different perspectives for evaluating and managing AI risk. Because many stakeholders take part in developing and deploying AI systems, the first approach taken is a lifecycle approach. For each lifecycle step, the decisive actors can be included and managed to ensure a risk-averse system. The first main category is the **application context**, in which systems are planned and designed but also deployed and monitored. The second stage is the **Data and Input** stage, in which the needed data is collected and processed. The third is the **AI model** stage, in which the model is built and used but also verified and validated. **Task and output** is the final stage in which the model is deployed in the production environment [18]. An insightful visualization of the lifecycle steps, activities necessary during each stage and involved actors can be found in Fig. 1.

The framework gives a definition of risk, being the likelihood and magnitude of an event. The magnitude would be defined by the negative impact of an event or the harm caused by an event [18]. This definition is used to look at societal values and see which could be impacted by typical AI risks. Those societal values will be in turn used to evaluate each lifecycle step on its conformity. The different values can be found in Fig. 2. Because most other frameworks implement a comparable value-based approach, the AI RMF also adds a mapping of its defined values with other suggestions, which can be found in Fig. 7. The other sources are an OECD working paper, which will be discussed in the following, the *EU AI Act*, which we mentioned before, and an executive order from the US. Before talking more about the origin of the framework, we will give a short definition of the value categories, as provided by the AI RMF.

Lifecycle	Activities	Representative Actors
Plan & design	Articulate and document the system's concept and objectives, underlying assumptions, context and requirements.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Collect & process data	Data collection & Processing: gather, validate, and clean data and document the metadata and characteristics of the dataset.	Data scientists, domain experts, socio-cultural analysts, human factors experts, data engineers, data providers, TEVV experts.
Build & use model	Create or select, train models or algorithms.	Modelers, model engineers, data scientists, developers, and domain experts. With consultation of socio-cultural analysts familiar with the application context, TEVV experts.
Verify & validate	Verify & validate, calibrate, and interpret model output.	
Deploy	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	System integrators, developers, systems/software engineers, domain experts, procurement experts, third-party suppliers with consultation of human factors experts, socio-cultural analysts, and governance experts, TEVV experts, end-users.
Operate & monitor	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives and ethical considerations.	System operators, end-users, domain experts, AI designers, impact assessors, TEVV experts, product managers, compliance experts, auditors, governance experts, organizational management, end-users, affected individuals/communities, evaluators.
Use or impacted by	Use system/technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.	End-users, affected individuals/communities, general public; policy makers, standards organizations, trade associations, advocacy groups, environmental groups, civil society organizations, researchers.

Fig. 1. Lifecycle steps, activities and actors [18]

Fig. 2. Value categories from the AI RMF

(1) Valid and Reliable: AI system validity and trustworthiness rely on accuracy and robustness. Inaccurate or unreliable AI systems that can't handle different data are risky and untrustworthy. Accuracy, reliability, and robustness should be measured, considering both computational aspects and human collaboration. These measurements must come with clear test details and be part of the system's documentation. Validity and reliability are checked through ongoing audits to ensure intended performance and minimize harm.

(2) Safe: AI systems must not cause harm to humans, their well-being, or the environment. Safe AI requires responsible design, clear usage guidelines, and cautious decisions by operators and users. Safety can be ensured by considering it in planning, using simulations, real-time monitoring, and the ability to modify or shut down systems going astray. AI safety lessons from fields like transportation and healthcare apply.

(3) Fair and Bias Is Managed: Fair AI aims at equality, addressing bias and discrimination. Fairness varies culturally and contextually. Bias comes in systemic, computational, and human forms. It's crucial to manage these biases,

even without explicit prejudice. Biases can be hidden in AI data and processes, affecting decisions and perpetuating harm. Transparency and fairness are linked.

(4) Secure and Resilient: Resilient AI systems handle unexpected changes, while secure systems protect data and functions. Resilience is about recovery, security is about prevention. Both relate to robustness but go beyond. Security issues include attacks and data breaches.

(5) Transparent and Accountable: Transparency involves revealing AI system details to users. It doesn't guarantee accuracy or other qualities. Accountability involves responsible parties in case of issues. Shared responsibility among AI actors is important. Cultural and legal contexts influence risk and accountability.

(6) Explainable and Interpretable: Explainability shows how an AI works, interpretability explains its output's meaning. Both aid in effective and responsible AI operations. Explaining models for different user knowledge levels enhances explainability. These systems are easier to monitor and audit.

(7) Privacy-Enhanced: Privacy safeguards autonomy and identity. Norms like anonymity and control should guide AI system design. Privacy risks overlap with security, bias, and transparency. Technical features can enhance or reduce privacy, and data processing's impact on privacy should be assessed.

For this framework, two working papers published by the *Organization for Economic Co-operation and Development* (OECD) are used as a blueprint. Those are the *OECD Framework for the Classification of AI Systems*, which in many ways is similar to the risk classification mentioned for the *EU AI Act* [3]. And the *Recommendation of the Council on Artificial Intelligence*, which in turn is quite similar in its approach to the value-based perspective just described for the NIST framework [4]. In Fig. 3 you can see the timeline in which the AI RMF was developed. Overall it is an iterative approach. Because all this development was still preceded by some additional ISO standards and cybersecurity research and is also being followed up by more publications, we decided against trying to create a complete timeline.

One important comment for this and other frameworks is, that we are leaving out organizational dimensions, that are also important for those frameworks. The AI RMF defines the dimensions *Govern, Map, Measure* and *Manage*, which are all needed to implement and use many AI security steps in a big organization. With this organizational perspective being outside the scope of this research, it would be very interesting for further evaluation and comparison.

2.3 Google

This leads us to the *Secure AI Framework* (SAIF), just published recently by Google. According to Google, the NISTs' AI RMF was taken as a starting point for their work [6]. Besides the obvious fact, that this is a lot easier than starting from scratch, it is a decisive step to get more practitioners and entities to start using Google's SAFE framework, by just being an evolution of the NIST framework.

Defining six crucial aspects to consider when developing secure AI systems. Those are intertwined with lifecycle approaches and also some value perspectives. Because those have been mentioned before, we will focus on the new aspects [8]. Google emphasizes the categories as being non-sequential and far from being a complete blueprint for

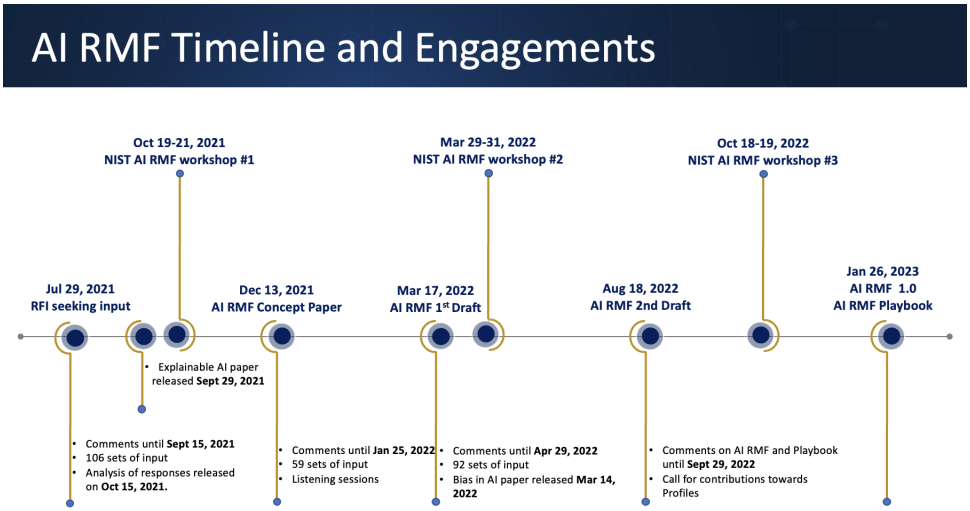


Fig. 3. AI RMF Timeline [18]

developing safe AI systems. The headers are taken directly from Google's publication, the descriptions are shortened and partly rephrased [8].

(1) Expand strong security foundations to the AI ecosystem: To use and expand robust cybersecurity systems for AI instead of developing from scratch. Expand where AI-specific security makes sense.
(2) Extend detection and response to bring AI into an organization's threat universe: Monitor inputs and outputs of systems and deploy AI-specific threat detection.
(3) Automate defences to keep pace with existing and new threats: Not just to protect AI against attacks, but also to deploy AI to scan for and prevent attacks, more research is needed.
(4) Harmonize platform-level controls to ensure consistent security across the organization: Consistency helps to mitigate and scale protections across different platforms and tools to ensure the best protection for all AI applications in a scalable and cost-efficient manner. This also means building controls and protection into the software development lifecycle.
(5) Adapt controls to adjust mitigations and create faster feedback loops for AI deployment: Constant testing can ensure detection and protection capabilities address the changing threat environment. This includes techniques like reinforcement learning based on incidents and user feedback and involves steps such as updating training data sets, fine-tuning models to respond strategically to attacks, and allowing the software that is used to build models to embed further security in context.
(6) Contextualize AI system risks in surrounding business processes: Conducting end-to-end risk assessments can help inform decisions. This includes an assessment of the end-to-end business risk, such as data lineage, validation and operational behaviour monitoring for certain types of applications. Organizations should construct automated checks to validate AI performance.

Additionally, Google defines different approaches from which to evaluate the overall security of an AI system. The main categories mentioned are Security, AI/ML model risk management, Privacy and compliance, and People and

organization [7]. Those categories are somewhat comparable to previously mentioned values, albeit summarizing topics into one category. New is the category of *People and organization*. The explanation given here is the need for governance structures and to manage the talent gap inside an organization.

Unrelated to the main framework, Google also published a paper taking the perspective of the extreme risk of AI. The publication *Model evaluation for extreme risk* discusses different ways in which AI systems can be used maliciously. The authors distinguish between a model's capability of causing harm and its willingness or propensity of doing so, and how this alignment question could be measured [25]. While focusing on those topics, they intentionally leave out the structural risk of models and incompetency issues, as they depend much more on the external world. Besides the usual lifecycle perspective, they propose the idea to focus on evaluating new capabilities of models. Decreasing the scope of model evaluations and allowing a more thorough analysis. They also call for external audits, regulation, systematic tracking, and evaluation of AI models by policymakers.

Finally, they also published a framework on how to implement internal audits [23]. This, however, seems to be out of the scope of this research.

2.4 Microsoft

In their recently published white paper *Governing AI: A Blueprint for the Future*, Microsoft states their commitment to implementing NIST's AI RMF, but also furthering their internal efforts to ensure the responsible creation and use of AI systems [16]. They give a high-level overview of the steps they are taking as an organization to be ready to responsibly develop AI. In their effort to operationalize those high-level goals, they set out to define multiple dimensions, that in many ways are in line with what we previously called values [15]. The dimensions are Fairness, Reliability & Safety, Privacy & Security, Transparency, and Accountability. Each topic contains subtopics, which often take a lifecycle perspective. With many of the aspects already having been presented thoroughly, we will not repeat them extensively.

2.5 Journal of Physics

The article *An Artificial Intelligence Security Framework*, published in the Journal of Physics [9], is the only framework found with no direct ties to companies or government entities, possibly making it the framework with the least bias. They describe a lifecycle approach, taken from an ISO standard, which is in line with the mentioned NIST framework. They offer an insightful visualization, which tries to merge all aspects into Fig. 4. Describing the previously mentioned values as *AI Security Goals* [9]. The goals define how AI systems should operate, and the different capability grades ensure the achievement of those goals from different perspectives, all while being supported by *AI Security Technology* and *AI Security Management*. Not as broadly used is the *Graded Capability of AI Security*. Covering Architecture Security, Passive defense, Active defense, Threat Intelligence and Offense. The closest match to the other frameworks would be the six-step approach by Google.

After this thorough introduction to some of the most relevant AI security and privacy frameworks, we will now present the mapping process we implemented to explore the scientific literature systematically. We will also present the research questions and discuss the research scope we deem fit for the topic and the course setting.

3 RESEARCH METHODOLOGY

Based on our research question and the related work chapter, we will introduce the research methodology in this chapter. The results and discussion will be built upon the outcomes of this methodology section.

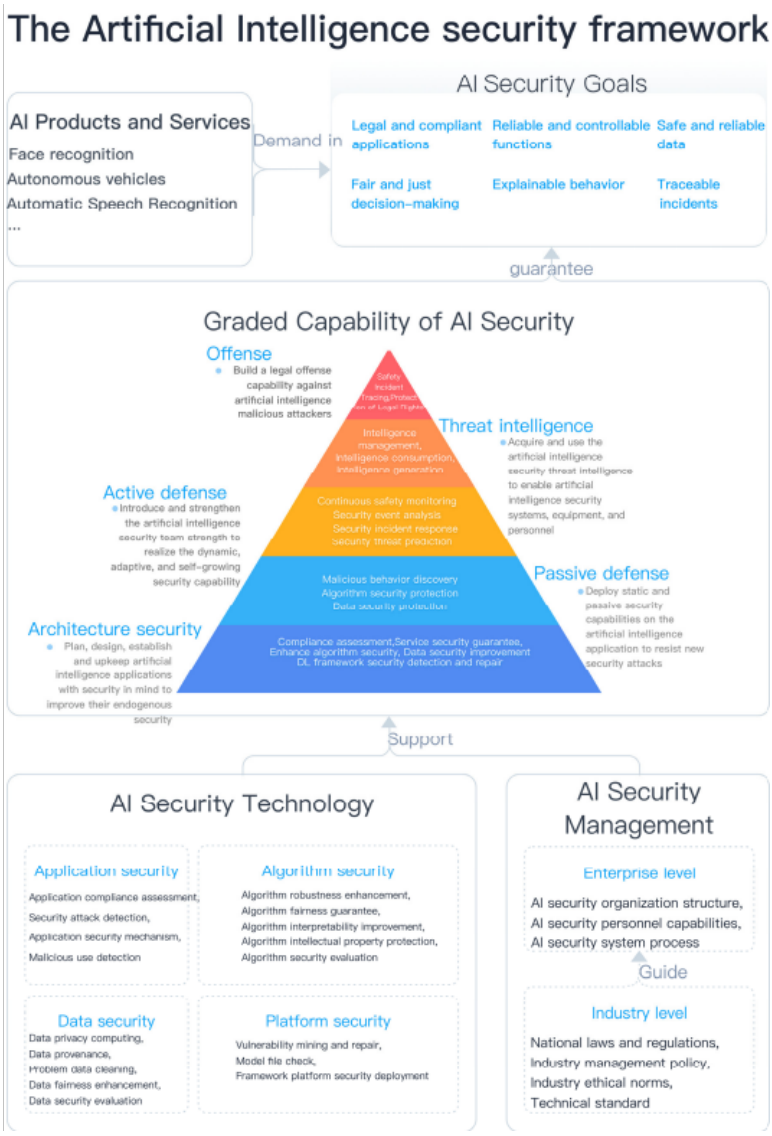


Fig. 4. Model of the *[An] Artificial Intelligence Security Framework* [9]

3.1 A systematic mapping process

We are using the systematic mapping process according to Petersen et al. (2008) to plan and conduct our literature search. We also considered conducting a systematic literature review. However, since systematic mapping is specifically tailored to the software engineering or IT domain, we have chosen to use this procedure. As an additional argument for the systematic mapping, we want to adopt some approaches from Pahl et al. (2017), as they offer interesting perspectives for our discussion. This approach helps to systematically search and classify literature based on methods to analyze and document results [20].

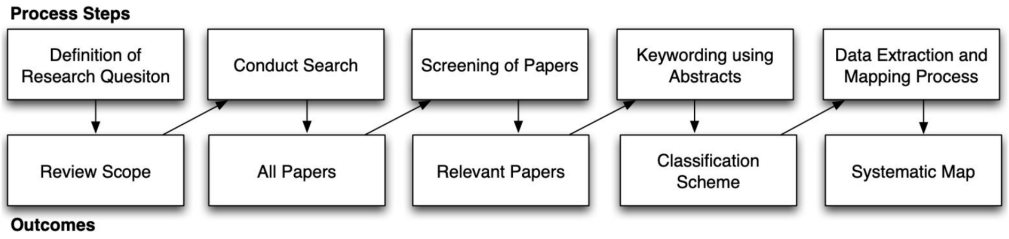


Fig. 5. Systematic Mapping Process from Peterson et al. (2008)

In Fig. 5 we can see a diagram from Peterson et al (2008), which illustrates their work process briefly. First of all, it is crucial to establish the research question and overall scope. Following this, the search is performed for multiple databases using a defined query. Subsequently, the papers are reviewed, and relevant ones are chosen based on inclusion and exclusion criteria. Afterwards, abstracts of the relevant papers are used for keywording. This aids in creating a classification scheme for data extraction. Finally, they create a systematic map based on the information they have gathered from the papers. We used this process as a guideline, the following paragraph describes the specific steps we want to discuss.

First, we defined our research scope, and the research question in 3.2 and selected the relevant databases in chapter 3.3. Next, we created the search strings and based our query on them in the following chapter. Then we applied our inclusion and exclusion criteria to the results in section 3.5. Finally, in the last chapters, we define the categories for the systematic map to perform the keywording of the abstracts and realize the data extraction from the selected papers to introduce the systematic map.

3.2 Research Question and Research Scope

We initially planned to work on the topic of “Security of AI Systems”. Based on our initial research, without a systematic approach yet, various surveys in the literature have approached the security of machine learning and AI systems from distinct perspectives. Those are the initial security aspects of any software product, not related to AI, algorithmic security aspects on multiple levels, data security aspects and where adversary attacks can be expected in those systems. Research focused on adversary attacks mostly analyses classification models during training and testing phases. Other research does not focus on all possible attack types but more on MLOps and general product lifecycle best practices for guaranteeing the best possible security of AI software. We took all aspects into account and came up with conducting a systematic mapping on the following research question:

„What are the key frameworks used to manage the technical and societal implications of Artificial Intelligence technologies and are they in line with recent developments in AI security and privacy research?“

The essential goal of a systematic mapping process is to provide a content overview of the research area and to identify the essential keywords [22]. It is also the research scope to identify the quantitative development of publications in the field. Furthermore, we want to identify what type of evaluation method, contribution type and novelty they represent. Finally, we will create categories to classify the keywords. A systematic approach is followed in all steps to make the results understandable and reproducible.

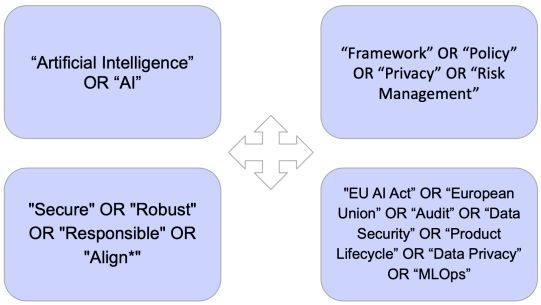


Fig. 6. Keywords for the query

3.3 Relevant Databases

To find the papers for our systematic mapping we needed to choose the appropriate databases. Our aim was to choose the main databases for high-quality papers in the field of computer science and software engineering. We have followed the choice from Petersen et al. (2008) and also selected IEEExplore and ACM Digital Library.
The IEEE database has a very high reputation in the field of computer science and software engineering. This database additionally offers a wide range of research papers, especially journals and very recent conference papers. By using these resources available through IEEE, we aimed to access the latest advancements and insights in our research domain.
In addition to IEEE, we also included the ACM database in our search strategy. ACM provides an interdisciplinary perspective by hosting contributions from various fields.
Lastly, we incorporated the Ebsco Host database into our research process. With EbscoHost we have access to seven out of eight journals in the “Basket of Eight”, compiled by the Association for Information Systems. These journals are widely regarded as leading publications in the information systems research field.
This way we intended to cover a broad range of journals for our systematic mapping process. Science Direct did not accept our query due to it being too extensive. With Peterson et al. (2008) emphasizing the importance of consistency and scientific rigour, we did not feel comfortable including Science Direct with an altered query. This will be discussed in more depth in the next chapter.

3.4 Search Strings and Query

For our search process, we initially employed broad search strings to capture a wide range of articles related to AI in combination with aspects such as framework, policy, privacy, and risk management. This approach yielded a substantial number of results that show the relevance of our research topic. However, to ensure the precision and relevance of our findings, we recognized the need to refine our search query. The final query can be seen in 6.

To refine our query, we took into account several factors. First, we considered the insights gained from the initial search and related work chapter. Furthermore, we looked at the keywords in the abstracts of interesting papers identified during the initial search iterations. By analyzing the abstracts of these papers, we gained a deeper understanding of the terminologies commonly used in the field. These insights allowed us to refine our search strings and create a more targeted query. As a last keyword, we filtered for “Framework” to exclude papers that do not refer to AI security frameworks. All of the boxes are connected with an “AND”-clause in the query, this yielded 95 relevant papers.
Unfortunately, we were unable to utilize Science Direct as a database due to the fact that our query involved significantly

more than eight connectors. Such a query cannot be executed using the Science Direct advanced search feature. Because the same query for each database is an integral part of the systematic mapping process to get reproducible results, we were unable to make use of this database. We attempted to omit certain keywords, but this would have negatively impacted the results for other databases as well if we had simplified the query for all databases. Additionally, we tried splitting the query into two parts. However, each part yielded output in the four to five-digit range, and it was not feasible to identify the intersection of papers present in both sets of results.

3.5 Inclusion and Exclusion Criteria

In the next phase of our research process, we applied inclusion and exclusion criteria to the final set of search results to exclude papers that were not relevant to answering our research questions [22].
First, we assessed the relevance of each article to our research questions. While most articles were directly related to our topic, we encountered for example one paper that focused on the historical perspective of the European Union, so we excluded it from our final selection. We reflected on adjusting the query because we got a paper that was not even referring to our field of studies but decided against it because it was an exception. Another criterion we considered was recency. Although the majority of the articles are recent, we excluded one paper from 1994 that was outdated and not relevant to the current context of our study.
Regarding the forum type, we included articles from various sources such as journals, magazines, books, and early-access publications. However, we excluded a few news articles as they typically lack the depth for our research purposes.
Language was another important criterion for exclusion. While our primary language is English, we were open to considering articles in other languages that we are proficient in. We came across a Croatian paper that did not meet our language criterion and was therefore excluded from our final selection.
Additionally, we ensured the exclusion of duplicate articles to maintain the uniqueness of our research dataset.
Lastly, we decided to exclude papers that primarily focused on cybersecurity, as our research revolves around AI frameworks, privacy, and risk management.

3.6 Classification Scheme

Following the approach by Petersen et al. (2008), we conducted keywording to implement a classification scheme based on the relevant papers. We used keywords from the formal methodology aspects presented in the paper by Pahl et al. (2017), keywords we identified by reading the abstracts, and primarily from the concepts derived from related work.

To conduct our mapping systematically, we used an analytical perspective given by the *EU AI Act* and the frameworks described in the related work. As all frameworks take a lifecycle approach, this was our initial perspective for structuring all research. This will be done in an exhaustive approach with the intent to be as precise as possible in the first mapping phases. Depending on the outcome, the categories might be consolidated in later parts of this research to derive insightful mappings. Multiple categories can be selected for each paper. This also applies to the value-based approach, which will be defined after.

3.6.1 Lifecycle approach. All papers taking a full lifecycle approach will be classified as a lifecycle paper and not be classified in other categories as well. For all other papers, the categories in Table 1 are derived from the NIST framework but also extended by categories mentioned in other approaches.

Table 1. Lifecycle steps and descriptions

Lifecycle Step	Description
Inception	Initial planing phase: what data is needed, what is the model supposed to achieve, what infrastructure is needed
Data collection	Collecting data or setting up structures for automated collection
Data cleaning	Make sure data is as expected, remove unneeded data, check for data injection
Data processing	Turn the data into a format usable by AI models, normalization, vectorization
Model design	Select or design model, input and output structures
Model training	Train the model with the previously selected data
Model testing	Test performance of model on specific tasks
Optimization	Do optimization for inference time, accuracy or use-cast specific metrics
Model evaluation	Final evaluation if model accuracy is sufficient, model is aligned and fit for purpose
Deployment	Put AI systems in production environment and make it accessible for inference by costumers
Monitoring	Monitor inputs and outputs of models, performance metrics, security matrices, alignment
Re-training	Re-train models once performance decreases or more data is available
Retirement	Remove model from production environment, delete data with privacy concerns

Table 2. A collection of value keywords and our final mapping keyword

Summary of keywords	Our mapping key used
valid, reliable, (technical) robustness, accurate, monitored	robustness
safety	safety
fair, bias-managed, non-discrimination, diversity, data governance, alignment	bias
secure, sequrity, resilient, resilience	techical security
transparancy, accountable, responsible, human agency, tracable, lawful	alignment
explainable, interpretable, understandable	explainability
privacy-enhanced, human rights, data governance, privacy, lawful	privacy

3.6.2 Value approach. The second mapping approach is the value-based one. This approach has been used multiple times. We are focusing on the AI RMF implementation provided by NIST. They already do an insightful mapping of their categories to the ones used in the *EU AI Act* and other frameworks.

3.6.3 Methodology Mapping. For the methodology of the papers, we examined the following aspects. First, we investigated the publication format and contribution type based on the paper by Pahl et al. (2017). Additionally, we referred to the definitions provided in the Research Type Facet by Petersen et al. (2008) to determine the contribution type for our final set of results. With these aspects, we can cover a broad spectrum to categorize the paper methodologies. In the following chapter, the results of all categories will be evaluated graphically, showcasing which categories were the most prevalent.

AI RMF	OECD AI Recommendation	EU AI Act (Proposed)	EO 13960
Valid and reliable	Robustness	Technical robustness	Purposeful and performance driven Accurate, reliable, and effective Regularly monitored
Safe	Safety	Safety	Safe
Fair and bias is managed	Human-centered values and fairness	Non-discrimination Diversity and fairness Data governance	Lawful and respectful of our Nation's values
Secure and resilient	Security	Security & resilience	Secure and resilient
Transparent and accountable	Transparency and responsible disclosure Accountability	Transparency Accountability Human agency and oversight	Transparent Accountable Lawful and respectful of our Nation's values Responsible and traceable Regularly monitored
Explainable and interpretable	Explainability		Understandable by subject matter experts, users, and others, as appropriate
Privacy-enhanced	Human values; Respect for human rights	Privacy Data governance	Lawful and respectful of our Nation's values

Fig. 7. Mapping of value approaches [18]

3.7 Systematic Map

Based on the process outlined in this whole chapter and the classification scheme described in the previous section, which includes all categories and potential sub-categories, we were able to create the systematic map. We implemented two versions of the systematic map. The first version can be found in the appendix, divided into two parts 20 and 21. The current and second version is built upon the initial version, consolidating categories for better interpretability. The subsequent discussion will provide a more detailed explanation of how the process of merging terms was conducted by us. To ensure readability, it was also divided into two graphics 8 and 9. The columns represent the defined lifecycle steps, values, and methodology aspects. An "X" was marked for each category a paper addressed. For the methodology categories, we used abbreviations explained in Table 10. The rows are numbered consecutively for all the papers we examined. The corresponding references can be found in the appendix under various tables 16, 17, 18 and 19. There, complete references for each paper are provided for further investigations. The following results and discussion are largely based on this systematic map and the process just described.

4 RESULTS AND VISUALIZATION

Our systematic mapping study yields a variety of insightful observations. Petersen et al. (2008) emphasize in their conclusion the significance of presenting outcomes through visual representations, a practice that should be adopted

	Lifecycle						Value-Based							Methodology		
	AI System Architecture	Data Management	Model Training	Deploy,Predict & Monitor	Retirement	Complete Lifecycle	Robustness	Safety	Bias	Technical Security	Alignment	Explainability	Privacy	Publication Format	Contribution Type	Evaluation Method
1	X									X				C	P	Ep
2			X				X						X	C	S	Ep
3														J	S	Ep
4		X					X					X	X	J	S	Ex
5		X	X										X	C	S	Ex
6			X							X			X	J	S	Ep
7													X	J	S	Ep
8		X	X										X	J	S	Ep
9	X	X								X				J	E	LR
10		X	X				X					X	X	J	S	Ep
11			X				X							C	S	Ep
12				X			X			X		X		J	S	Ep
13			X							X				C	S	Ex
14						X	X							C	S	CS
15	X		X	X									X	J	S	Ep
16	X		X	X			X			X			X	J	S	Ep
17	X	X		X			X			X			X	J	S	Ep
18		X	X	X			X			X			X	C	S	Ex
19						X		X	X		X			C	S	Ep
20			X							X			X	J	S	Ep
21	X		X		X		X			X				C	S	Ep
22										X			X	J	E	LR
23	X			X						X				C	EP	CS
24							X			X				J	E	LR
25			X	X			X			X			X	C	E	Ex
26													X	C	S	Ep
27			X				X						X	C	S	Ep
28												X	X	J	E	Ex
29	X		X				X			X			X	J	S	Ep
30	X		X				X						X	J	S	Ep
31	X	X	X							X			X	J	S	Ex
32	X		X	X						X			X	C	S	Ep
33	X									X				C	S	Ep
34	X						X			X			X	C	S	Ep
35	X		X	X									X	J	S	Ep
36						X				X			X	J	S	LR
37	X		X							X				J	SE	LR
38	X		X	X			X						X	J	S	Ep
39	X	X	X	X			X			X			X	C	S	Ex
40										X			X	C	S	Ep

Fig. 8. Systematic Map - Table (Part 1)

more widely in systematic reviews. As a result, we will present some fundamental findings, visualize them where we see fit, and elucidate the insights we have gained. Additionally, we will provide an overview of the most recent topics in AI security and privacy that we have extracted from the chosen literature.

	Lifecycle						Value-Based							Methodology		
	AI System Architecture	Data Management	Model Training	Deploy,Predict & Monitor	Retirement	Complete Lifecycle	Robustness	Safety	Bias	Technical Security	Alignment	Explainability	Privacy	Publication Format	Contribution Type	Evaluation Method
41	X			X			X			X			X	J	S	Ep
42			X											C	S	Ep
43			X	X			X			X			X	C	S	Ep
44		X	X				X			X			X	J	S	Ep
45			X				X			X				J	E	Ep
46	X	X	X				X	X				X	X	J	E	LR
47							X	X		X				C	S	Ex
48	X	X		X			X			X			X	C	S	Ex
49	X			X		X				X				J	E	M
50						X	X			X			X	J	S	Ep
51	X			X									X	C	S	Ep
52			X							X			X	C	S	Ep
53				X			X	X		X				C	S	Ep
54	X		X	X			X			X			X	C	S	Ep
55	X		X				X			X				C	S	Ep
56		X	X				X			X			X	C	S	Ex
57			X										X	J	S	Ep
58	X		X	X			X	X		X				J	S	Ep
59													X	J	S	Ep
60	X			X			X	X		X				C	S	Ex
61		X	X										X	C	S	Ep
62		X					X	X		X				J	S	Ep
63	X			X			X			X			X	C	S	Ex
64						X	X	X			X			C	S	LR
65		X								X	X		X	C	S	Ex
66										X			X	J	E	CS
67										X			X	C	S	CS
68						X		X	X		X	X		B	S	Ex
69	X			X			X	X		X				C	S	Ep
70		X											X	J	S	Ep
71			X				X			X			X	J	S	Ep
72	X	X	X	X			X			X			X	C	S	Ex
73										X			X	C	S	Ep
74	X			X			X	X		X	X			C	S	Ep
75	X		X	X			X			X			X	J	S	Ep

Fig. 9. Systematic Map - Table (Part 2)

A more in-depth discussion of the framework perspectives and what we were able to learn, will be done in the discussion part of this research.

4.1 Temporal overview of studies

AI security is an emerging topic in literature. By leveraging AI in different domains, AI security is gaining importance and is drawing attention both in academia and industry.

Methodology Categories		
Publication Format	Journal	J
	Conference	C
	Book	B
Contribution Type	Evaluation	E
	Solution	S
	Philosophical	P
Evaluation Method	Experiment	Ep
	Example	Ex
	Case Study	CS
	Meta Study	M
	Literature Review	LR

Fig. 10. Systematic Map - Abbreviations

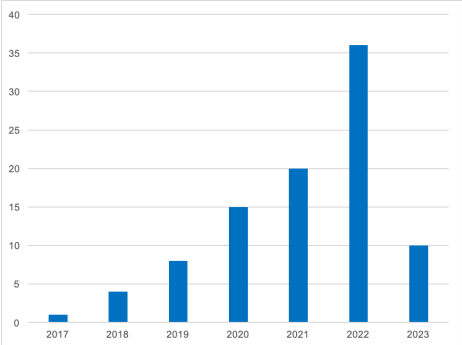


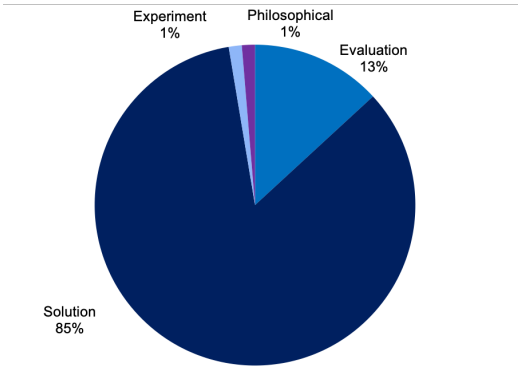
Fig. 11. Study Distribution by Year

As shown in Fig. 11, by the time of writing, the number of published papers has an increasing trend from 2017 to 2022. However, the number of published papers in 2023 is less than that of 2022. This is mostly due to proceeding papers to be published later in the current year. As a result, the list of papers for 2023 is inevitably left incomplete. The data demonstrates the growing significance of AI security and privacy within the area of computer science research. Further increase in the future is to be expected. This underlines the significance of our paper. It is essential to maintain an overview in a rapidly expanding research field and systematically map the key aspects, particularly for future research or frameworks.

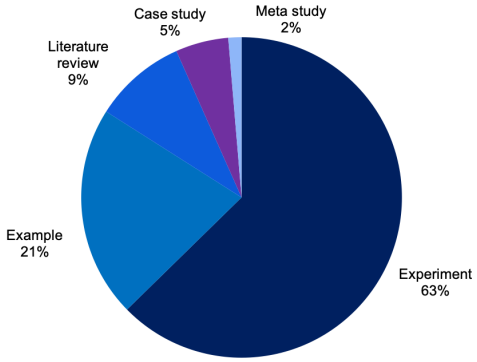
4.2 Research methods

In Fig. 12a, the primary studies are depicted according to their **contribution type**, as defined by Peterson et al. (2008). The types consist of solution, evaluation, experiment, and philosophical paper. 85% of the studies introduce innovative or substantially enhanced techniques along with supporting reasoning. They are classified as the contribution type solution. 13% percent evaluate existing solutions or explain a solution from a practical point of view, Thus are classified as evaluations. One percent is classified as primarily an experiment and one percent is primarily a philosophical paper. The high percentage of the contribution type solution again speaks for the recent rise of the topic of AI security and privacy.

In 12b, we illustrate the **evaluation methods** used by researchers. Again, using the categories provided by Peterson et al. (2008). Experiments are the most common method with 63% and examples are the second most with 21%. Literature review, case study and meta-study can be found with nine percent, five percent and two percent respectively. Again, this especially shows how new the topic is. In some time, we expect case studies, meta-study and literature reviews to rise, while experiments and examples might decline percentage-wise over time.



(a) Study Distribution by Contribution Type



(b) Study Distribution by Evaluation Methods

4.3 Overview of studies - AI lifecycle stages

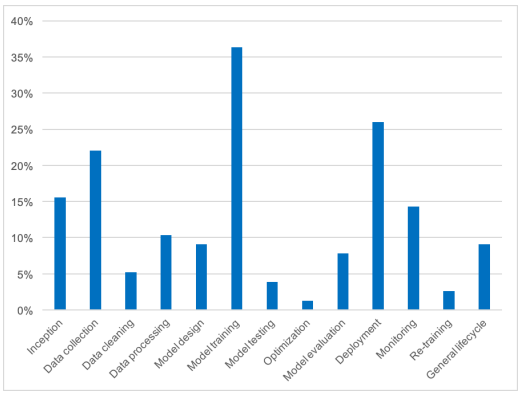
In previous sections, we discussed the lifecycle of AI regarding security and privacy based on the AI RMF and other frameworks. In this section, we classify the papers based on the stage of the lifecycle they focus on. It should be taken into account, that many papers address more than one stage of the AI lifecycle. We provide a bar chart under Fig. 13a, which illustrates the percentage of studies that include each stage of the AI lifecycle. For now, we look at all lifecycle steps. How they could be summarized will be presented in the discussion part. For now, data collection, model training, and model deployment are aspects of the AI lifecycle security that are addressed the most.

4.4 Overview of studies - Value mapping

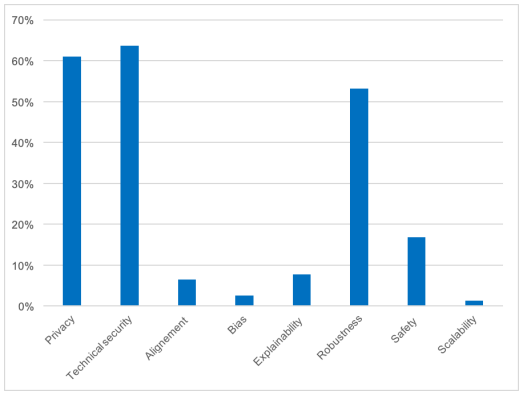
In this section, we focus on the value aspects of AI security. We provide a bar chart in Fig. (13b) that shows the percentage of studies with respect to the AI security values that they cover. Most of the studies address more than one value. Privacy, technical security, and robustness can be observed the most in the underlying literature. With bias, alignment, explainability and safety still being discussed in depth in the presented frameworks and also the *EU AI Act*, we cannot argue for them being less important. Two assumptions can be derived from this observation. On the one hand, our query can play a crucial role in the distribution of the value occurrences, on the other hand, alignment has just recently arrived in the generally used vocabulary since the widespread adoption of ChatGPT. This overall shows the need for more exploratory research with a focus on bias, alignment, safety and explainability.

4.5 Publication formats

Regarding the sources of studies, we recognized the following publication formats: papers for journals and magazines, conferences, and books. In Fig. 14 we can see that almost all resource formats are conferences and papers. This seems



(a) Percentage of studies based on Each Stage of AI lifecycle



(b) Percentage of studies Addressing AI Security Values

to be the common practice in computer science or cybersecurity research. The substantial count of conference papers as a publication format underscores the role of conferences in facilitating knowledge exchange and sharing the latest research advancements in the AI security domain.

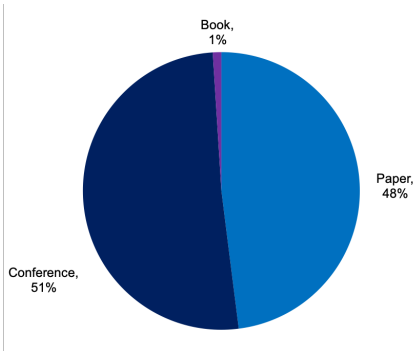


Fig. 14. Study Distribution by Publication Format

5 DISCUSSION: AI SECURITY AND PRIVACY

5.1 Recent topics in AI Security and Privacy research

Overall, there are some recurring topics that can be found in the literature. We will give an introductory overview of how the topics are related and interdependent. We will also look at, in which lifecycle step the topics are most likely to be found, and what value-based goals can be achieved or supported with each topic.

The most prominent topic is federated learning, which could be subdivided into multiple other topics. One key idea of federated learning is to keep sensitive data on the end consumer device. This leads to privacy-enhanced learning. This approach can also be merged with blockchain or distributed ledger technology, to either enable privacy-enhanced centralized training or create practical consensus algorithms for federated learning and inference or improve the efficiency of training. Those aspects are especially useful for IoT applications and edge computing.

Some more independent topics are adversarial attacks and defences against those. One form of defensive mechanism can be the application of AI. Finally, a topic just noted in two papers is model liability, ownership, and copyright. Before discussing those topics in depth, we will first summarise the lifecycle steps comprehensively.

5.2 Deriving relevant lifecycle steps from the prevalent literature topics

After working through many computer science papers discussing AI security and privacy, we are more confident in consolidating the very in-depth lifecycle steps. We will argue for which categories can be merged in a sensible manner. Those decisions will be based on the topics found in the research papers and perspectives given in the frameworks. If multiple papers map to the same lifecycle steps, those steps will be considered to be merged. This is in many ways similar to the high-level categories proposed by other frameworks but also differs in some stages. Following this, a mapping of the resulting stages to the previously mentioned topics will be conducted. Most of the initial information concerning all the papers is depicted in our first version of the systematic mapping tables 20 and 21 in the appendix. The new version of the systematic map can be viewed at 8 and 9.

Many papers discussing federated learning, edge computing and privacy-enhanced learning have considerable requirements for model design and the backend infrastructure, making it necessary to incorporate those technical requirements already in the inception stage. This leads to the first suggestion to consolidate the inception stage, the Model design stage and the deployment stage into an overarching topic of *AI system architecture*. Data handling is a key aspect of federated learning architecture. However, due to its overall importance and stark difference, it will not be incorporated into this category.

When analysing the research literature for data topics, it is quite difficult to assign just one of the lifecycle steps to a paper. Data collection is especially looking at how to enable distributed training, privacy-enhanced training with distributed ledger technology or other implementations. The screening for maliciously injected data is the next important step. While checking the data quality will take place at the data cleaning step, it does not make sense for an organization to disconnect this step from the data collection. After data has been collected and, for instance, saved to a central company database, employees might expect internal datasets to be trustworthy, making the analysis for malicious data a necessary step of data collection.

There is research into noise injection during the data processing stage, as one defence against adversarial attacks on AI models [24]. However, this topic is not present in the systematically collected research, making it seem not as relevant from an overall AI security perspective. Because no other research can be found looking especially at data processing, it seems sensible to merge all three stages into an overall stage that could be named *Data Management*.

After the initial planning phase of the overall infrastructure, model architecture and the data handling phase, the next step is model training. In the reviewed literature, there is no clear distinction made between model training and testing, specifics for optimization are not mentioned whatsoever. Again, in most literature, no clear distinction is made between model testing and evaluation. Instead, when filtering for the value-based category of *alignment* or *bias*, we can find multiple papers discussing the topic of model evaluation. In the frameworks presented in related works, a clear distinction is made between testing models for technical measures like accuracy or loss and evaluating models on their alignment to societal values, fit for purpose and risk for unwanted societal impact. The initial draft of the AI RMF distinguishes those as technical characteristics and socio-technical ones [17]. Leading to two crucial lifecycle steps. One

being *Model training and validation* and the other being *Socio-technical Model assessment*.

While planning and requirement analysis for deployment are also part of the previously mentioned *AI system architecture*, the technical rollout, bug fixes and re-evaluation of the implemented backend are still necessary. The implementation of a robust monitoring system from a technical but also an alignment perspective should be finalized before any deployment steps are taken. Leading to the comprehensive lifecycle stage name of *deploy, predict and monitor*.

Re-training as a category will be excluded. It was rarely mentioned and theoretically consists of the stages *monitoring and inference* (when re-training is necessary), *Data Management* (collecting new data and processing it) and *model training and validation* with the new data, making the re-training category redundant.

While security and privacy aspects of retirement are rarely discussed, the importance of a thought-through retirement stage is mentioned multiple times in the literature. It should be best practice, to reflect on how to implement the system retirement in the *AI system architecture* and the *data management* stages. Because technical and regulatory circumstances can change, an extensive re-evaluation before retirement is important to ensure data privacy compliance and prevent model weights and sensitive data from leaving the ecosystem it was intended for.

Further investigation, into how far all those conclusions are dependent on the selected query for this research or the topic just not being present in the recent AI security literature is needed.

A mapping of the assigned lifecycle and value categories can be found in Fig. 15. We can see a strong correlation between robustness, technical security and privacy. *Model training* and *AI system architecture* have the highest count. Of course, it is important to consider that we are dealing with absolute numbers here. Overall, robustness, technical security, and privacy are frequently present in every lifecycle stage. Hence, it is also intriguing to examine the values with lower overall counts. For instance, safety appears predominantly in *AI System Infrastructure* and *Deploy, Predict Monitor*. Privacy is the most occurring category for *data management*, which is to be expected. Bias on the other hand is rather underrepresented.

5.3 In-depth discussion of relevant topics

5.3.1 AI privacy and security incorporated with IoT. The Internet of Things (IoT) and Artificial Intelligence (AI) are closely connected and often work together to create smarter, more efficient, and more capable systems. The combination of AI and IoT holds immense potential across various industries, such as healthcare [26]. There are 21 references which directly study this topic.

When discussing AI from a data management and model training perspective for IoT systems, federated learning is addressed multiple times and plays a significant role in proposed frameworks. Traditional machine learning models are limited to conducting training in a centralized manner, where data is gathered on a central server or dataset. This situation raises substantial security concerns since crucial data is transmitted between the data-generating entity and servers. However, in a *federated learning* (FL) approach, the model or data processing is distributed across the devices. Every device contributes to the creation of the global model by generating its own local model or weights and providing those to a central server [27]. In one of the studies, Khowaja et al. (2022) propose a federated learning and encryption-based private (FLEP) AI framework. This framework is designed for an Industrial IoT (IIoT) environment. The uniqueness of the proposed framework lies in private AI for data and model security, meaning that even the model and the practitioners never get to see the actual, unaugmented data. Regarding the private AI for data security, the

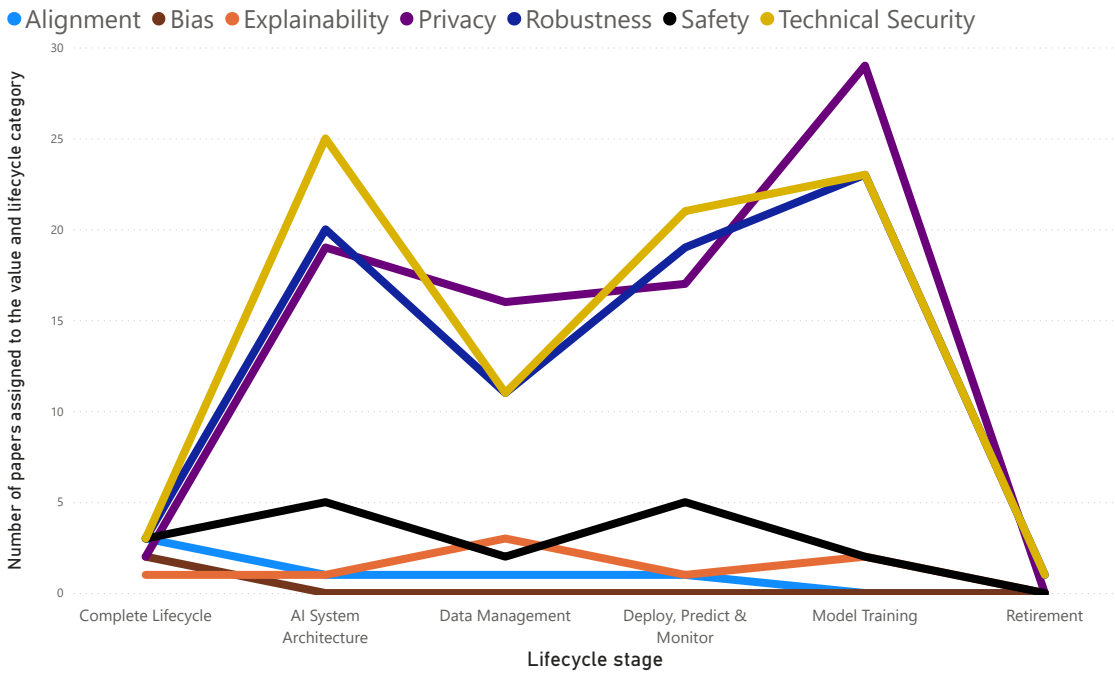


Fig. 15. Occurrence of literature representing specific values at every lifecycle step

FLEP AI framework allows any encryption method to be used on the data before sending it to a data analyst, but to show the module's realization a chaotic sequence and wavelet-transform-based data encryption method is proposed in the paper. The main focus is on the implementation of the method on image data. The private AI for the data security module can be divided into three segments: encryption, encoding, and adding noise. The keys used in generating chaotic sequences along with the secret image and noise parameters will be used to decode and decrypt the information at the data analysts' module in the framework.

Federated learning solves the issues regarding data ownership and governance to some extent because the data remains with the user [10]. However, using a standalone implementation does not guarantee model parameter security. FL does not store data but data can be reconstructed with attacks. When it comes to AI model security, the FLEP AI framework extends the security to model parameters as well. Homomorphic encryption (HE) techniques can secure model parameters generated at the data analysis phase. A data analyst generates paired keys with encrypted model parameters. The keys are shared with the trusted AI cloud and IIoT services for accessing the model parameters. If needed, the HE module sends the trained model parameters as plain text or encrypted form to the untrusted AI cloud for further computation without any paired key. The HE allows the untrusted AI cloud to perform operations on the encrypted text. The result of the operation will be sent to the trusted IIoT services.

On the other hand, focusing on data and model training privacy for healthcare AI, Cock et al. (2018) leverage techniques from secure multiparty computation (SMC) for private classification with tree ensembles. They implement

an SMC-based privacy-preserving machine learning algorithm on a proprietary healthcare analytics platform. Aiming to support physicians to drive better clinical outcomes in an accurate, scalable, and secure manner. Their proposed system allows two or more parties to jointly compute a specific output from their private information in a distributed fashion, without revealing the private information to each other [5]. The system can classify one party's input with another party's classifier. At the end of the interaction, the party with the classifier will not have learned anything about the other party's input, and the party with the input will not have learned anything about the other party's classifier.

AIoT as an integration of AI technologies with IoT infrastructure has drawn researchers' attention to itself too. In AIoT the IoT devices metaphor the digital nervous system and AI as the brain of a system [13]. The integration of AI enhances Edge computing, which is a decentralized computing approach, in order to provide a higher level of intelligence and smart capabilities to the IoT devices themselves. This combination of AI and Edge computing enables IoT devices to make intelligent decisions and perform advanced computations locally, without the need for constant communication with a centralized server or cloud.
Zheng et al. (2022) argue that although using federated learning and distributed machine learning has helped AIoT to realize its services by retaining the data locally on devices and only exchanging model parameters, existing IoT data trading methods fail to meet this secure and efficient process. Therefore, they introduced a framework for data trading over AIoT. According to their proposed framework, service providers can trade on deep learning model training instead of purchasing the full datasets. They address the design of the learning parameters and the comprehensive privacy concerns. The proposed framework includes how service providers can maintain the model's performance. They evaluate their proposed framework which results show model performance improvements [31]. In addition to FL, researchers have addressed Software-Defined Networking (SDN). SDN is a network architecture approach that enables the network to be intelligently and centrally controlled, or 'programmed,' using software applications.

High levels of data and model privacy have significant importance when it comes to healthcare systems. Reliable data transmission is the main requirement of a healthcare system. Therefore, many studies specifically focused on AI privacy in different stages in the healthcare industry. Otoum et al. (2022) focus on Healthcare 4.0. Healthcare 4.0 incorporates a broad spectrum of opportunities to enhance healthcare by leveraging Industry 4.0 technologies since it presents a novel and inventive perspective for the healthcare industry. With Health 4.0, it is anticipated that distributed and edge-supported AI will enable faster and more accurate early-stage disease discovery that relies significantly on intelligent remote and on-site IoT devices [19]. The mentioned paper proposes an FL-enabled framework for healthcare systems that is supported by edge-computing, blockchain and intelligent IoT devices. Comparing the FL supported model with pre-trained models, shows improvements with respect to data privacy and accuracy. Healthcare 5.0 employs AI and IoT as well and shifts from mass customization (Healthcare 4.0) to mass personalization. Studies on this system mainly discuss the challenges regarding data security.

5.3.2 Privacy-enhanced training. Overall 31 papers have been found to discuss the topic of privacy-enhanced training in some way or form. Yang et al. (2019) propose a federated learning solution that enables federated learning without sharing of data. Models can be trained decentralized and the trained algorithms can either be shared directly or via their weights. However, there is still some concern about backtracking from those weights and extracting meaningful private information [29]. An extension of this, albeit from a high-level perspective, is *knowledge federation*, the idea of sharing information learned from data on different abstraction levels. The local model outputs embeddings, which can be shared with other entities and interpreted with ensemble learning methods. This omits the issue of data privacy and

enables ensemble learning across entities [12].

As an alternative or add-on, blockchain or distributed ledger technology can be used in a distributed training context. It can be used to make robust and safe decisions in a trustless environment. Either concerning the distribution of computational load, especially useful in an IoT setting [14], or trustless, anonymous and secure consensus building, useful for ensemble learning with many small distributed models. Finally, blockchain technology can also be used to share private and proprietary data in a secure and trustless way, to enable multiple entities to train a shared model without giving other entities access to sensitive information [30].

5.3.3 Adversarial attacks. Papers discussing adversarial attacks on AI systems occurred just once in the scanned literature. AI is seen as one of the critical enablers of high-throughput 5G networks, making adversarial attacks on those systems a threat. The paper is assessing vulnerabilities inside those 5G networks towards those attacks [32]. This topic seems to be especially important in the *deploy, predict, monitor* lifecycle step.

5.3.4 Model ownership, copyright and liability. With the *EU AI Act* stating clearly, that the entity initially offering an AI service is also liable for malicious capabilities of those models, ownership and copyright of those models become increasingly important. Chen et al. (2022) propose a framework that measures the similarity of the input and output of two models, making it possible to derive assumptions if one model has been stolen [1]. This in turn shows, that it might be necessary for companies to save models in their retirement stage to defend against or make copyright claims in the future.

5.4 Geopolitical implications

With the mention in the related work, that this research is taking on a Eurocentric perspective, it should also be discussed, which countries or regions are furthering the research into AI security and privacy and therefore their interests as well. With the AI RMF and the publication of Google's and Microsoft's frameworks, a strong representation of U.S. interests can be found concerning this topic. On the other hand, many publishing researchers can be found to have Chinese names and a lot of research is published through Chinese universities or Chinese research institutes. Overall showing that, China and the U.S. seem to be at the forefront of AI research, while the EU, quite typically, is at the forefront of policy. Due to language and knowledge constraints, we were not able to search for complete AI frameworks provided by Chinese companies. Incorporating perspectives from e.g. Huawei, Tencent and Alibaba could generate some insightful learnings. From general knowledge, we can however say, that China would have very different value goals for evaluating AI.

6 CONCLUSION

Overall, we were able to give a meaningful overview of recent developments in AI security and privacy, on the policy and the framework side. The *EU AI Act* takes a big leap by deciding what AI systems society is allowing to be used and which systems are deemed to have too high a risk of negative impacts. By looking at the most impactful frameworks, we accomplished to derive key perspectives from each framework, but also how the framework's development is building on each other. We managed to show, that ISO and OECD took some important initial steps to create frameworks. Which in turn have been taken over by NIST, which is now being implemented and adapted by some of the biggest actors in the space of AI.

By mapping the lifecycle perspective of the frameworks to the systematically collected literature, we were able to present some of the recent developments in AI security and privacy and at which lifecycle step those research propositions can be implemented in a useful manner.

We were able to show, that AI security and privacy are gaining attention in recent years. Federated learning, federated consensus and data sharing via blockchain, adversarial attacks, IoT and copyright seem to be topics of interest lately.

Even though this was mentioned already in related work, we would like to emphasize again, that many organizational and management perspectives discussed in the frameworks did not get any meaningful attention in this research. It does not seem feasible for the scope of this research and additionally, is not the main focus of AI security and privacy research.

A potential shortcoming of this research is not finding any meaningful mapping to the retirement stage, one reason might be our selected query. Fundamentally, there is the issue that the selection of keywords and their assignment always introduces a certain bias. For example, it was noticeable that the keywords Technical Security and Privacy were associated with a large number of papers, although the overarching theme is, of course, AI Security.

Because the retirement stage is mentioned many times but not discussed in depth, we assume, that no specific research is present. A potentially interesting topic we found is how to save a model during retirement to still be able to make or defend copyright claims.

Further topics for which we expected more content in the systematically explored literature and see high potential for insightful research are more independent evaluations of the *EU AI Act*, the AI security and privacy frameworks, AI for defensive capabilities, and especially the emerging topic of alignment, fit for purpose and societal risks. All mentioned topics could benefit especially from interdisciplinary research. The paper discussing *Model evaluation for extreme risks* [25], many research papers leading up to the *EU AI Act* but also the AI RMF discuss risk, alignment and fit for purpose of models. A meta-study or a systematic literature study in this direction might also still be interesting for a more complete overview but also the distinction between Chinese, American and European perspectives.

Finally, we learned, that many parts of the regulation and the frameworks cover similar topics, not only from a technical perspective but also from a societal value perspective. This makes a lot of sense with most frameworks and regulations building on top of each other, citing each other and overall striving to create consensus over how AI should be developed and used. With most research focusing on sub-topics, especially proposing new solutions, full frameworks are rarely discussed on the research side of the explored literature. This can be taken as a sign of this field being very new overall.

The highest value add, that we see in this research, is informing practitioners on emerging technologies, where they can be applied in the AI lifecycle and what specific value goals can be achieved by doing so. Furthermore, we have provided a comprehensive overview of the central frameworks and the literature in an emerging field.

.

REFERENCES

[1] Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song. 2022. Copy, Right? A Testing Framework for Copyright Protection of Deep Learning Models. In *2022 IEEE Symposium on Security and Privacy (SP)*. 824–841. https://doi.org/10.1109/SP46214.2022.9833747

[2] European Comission. 2021. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. 206 final (2021).

[3] Organisation for Economic Co-operation and Development. 2022. OECD Framework for the Classification of AI Systems. (2022). https://www.oecd-ilibrary.org/docserver/cb6d9eca-en.pdf?expires=1692278075&id=id&accname=oid011384&checksum=BE9329432FB670EC594DE44732ECB480

[4] Organisation for Economic Co-operation and Development. 2022. Recommendation of the Council on Artificial Intelligence. (2022). https://legalinstruments.oecd.org/api/print?ids=648&lang=en

[5] Kyle Fritchman, Keerthanaa Saminathan, Rafael Dowsley, Tyler Hughes, Martine De Cock, Anderson Nascimento, and Ankur Teredesai. 2018. Privacy-Preserving Scoring of Tree Ensembles: A Novel Framework for AI in Healthcare. In *2018 IEEE International Conference on Big Data (Big Data)*. 2413–2422. https://doi.org/10.1109/BigData.2018.8622627

[6] Google. 2023. *Introducing Google's Secure AI Framework.* https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/

[7] Google. 2023. *Secure AI Framework Approach. A quick guide to implementing the Secure AI Framework (SAIF).* https://services.google.com/fh/files/blogs/google_secure_ai_framework_approach.pdf?hl=de

[8] Google. 2023. *Secure AI Framework (SAIF): A Conceptual Framework for Secure AI Systems.* https://developers.google.com/machine-learning/resources/saif?hl=en

[9] Huiyun Jing, Wei Wei, Chuan Zhou, and Xin He. 2021. An Artificial Intelligence Security Framework. *Journal of Physics: Conference Series* 1948, 1 (jun 2021), 012004. https://doi.org/10.1088/1742-6596/1948/1/012004

[10] Sunder Ali Khowaja, Kapal Dev, Nawab Muhammad Faseeh Qureshi, Parus Khuwaja, and Luca Foschini. 2022. Toward Industrial Private AI: A Two-Tier Framework for Data and Model Security. *IEEE Wireless Communications* 29, 2 (2022), 76–83. https://doi.org/10.1109/MWC.001.2100479

[11] Johann Laux, Sandra Wachter, and Brent Mittelstadt. [n. d.]. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance* n/a, n/a ([n. d.]). https://doi.org/10.1111/rego.12512 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/rego.12512

[12] Hongyu Li, Dan Meng, Hong Wang, and Xiaolin Li. 2020. Knowledge Federation: A Unified and Hierarchical Privacy-Preserving AI Framework. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*. 84–91. https://doi.org/10.1109/ICBK50248.2020.00022

[13] Junxia Li, Jinjin Cai, Fazlullah Khan, Ateeq Ur Rehman, Venki Balasubramaniam, Jiangfeng Sun, and P. Venu. 2020. A Secured Framework for SDN-Based Edge Computing in IoT-Enabled Healthcare System. *IEEE Access* 8 (2020), 135479–135490. https://doi.org/10.1109/ACCESS.2020.3011503

[14] Xi Lin, Jun Wu, Ali Kashif Bashir, Jianhua Li, Wu Yang, and Md. Jalil Piran. 2022. Blockchain-Based Incentive Energy-Knowledge Trading in IoT: Joint Power Transfer and AI Design. *IEEE Internet of Things Journal* 9, 16 (2022), 14685–14698. https://doi.org/10.1109/JIOT.2020.3024246

[15] Microsoft. 2022. *Microsoft Responsible AI Standard, v2.* https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

[16] Microsoft. 2023. *Governing AI: A Blueprint for the Future.* https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

[17] NIST. 2022. *AI Risk Management Framework: Initial Draft.* https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

[18] National Institute of Standards and Technology. 2022. AI Risk Management Framework: Second Draft. (2022). https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf

[19] Safa Otoum, Ismaeel Al Ridhawi, and Hussein Mouftah. 2022. Realizing Health 4.0 in Beyond 5G Networks. In *ICC 2022 - IEEE International Conference on Communications*. 2960–2965. https://doi.org/10.1109/ICC45855.2022.9838687

[20] Claus Pahl, Antonio Brogi, Jacopo Soldani, and Pooyan Jamshidi. 2017. Cloud Container Technologies: A State-of-the-Art Review. *IEEE Transactions on Cloud Computing* 7, 3 (2017), 677–692. https://doi.org/10.1109/TCC.2017.2702586

[21] Georgios Pavlidis. 2023. Deploying artificial intelligence for anti-money laundering and asset recovery: the dawn of a new era. *Journal of Money Laundering Control* 26, 7 (2023), 155 – 166. https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=163820805&site=ehost-live

[22] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic Mapping Studies in Software Engineering. *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering* 17 (06 2008).

[23] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. arXiv:2001.00973 [cs.CY]

[24] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. 2018. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness against Adversarial Attack. arXiv:1811.09310 [cs.LG]

[25] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, Lewis Ho, Divya Siddarth, Shahar Avin, Will Hawkins, Been Kim, Iason Gabriel, Vijay Bolina, Jack Clark, Yoshua Bengio, Paul Christiano, and Allan Dafoe. 2023. Model evaluation for extreme risks. arXiv:2305.15324 [cs.AI]

[26] Hemang Subramanian and Susmitha Subramanian. 2022. Improving Diagnosis Through Digital Pathology: Proof-of-Concept Implementation Using Smart Contracts and Decentralized File Storage. *J Med Internet Res* 24, 3 (28 Mar 2022), e34207. https://doi.org/10.2196/34207

[27] Ryhan Uddin and Sathish Kumar. 2022. SDN-based Federated Learning approach for Satellite-IoT Framework to Enhance Data Security and Privacy in Space Communication. In *2022 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE)*. 71–76. https://doi.org/10.1109/WiSEE49342.2022.9926943

[28] Michael Veale and Frederik J. Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act. *CoRR* abs/2107.03721 (2021). arXiv:2107.03721 https://arxiv.org/abs/2107.03721

[29] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* 10, 2, Article 12 (jan 2019), 19 pages. https://doi.org/10.1145/3298981

[30] Zheng Zhang, Liang Huang, Renzhong Tang, Tao Peng, Lihang Guo, and Xingwei Xiang. 2020. Industrial Blockchain of Things: A Solution for Trustless Industrial Data Sharing and Beyond. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. 1187–1192.

https://doi.org/10.1109/CASE48305.2020.9216817

[31] Xu Zheng, Lizong Zhang, Bei Hui, Ling Tian, and Zhipeng Cai. 2022. A Secure and Efficient Framework for Multi-Round Data Trading Over the Internet of Artificially Intelligent Things. *IEEE Internet of Things Magazine* 5, 1 (2022), 119–124. https://doi.org/10.1109/IOTM.001.2100194

[32] Mikhail Zolotukhin, Parsa Miraghaei, Di Zhang, and Timo Hämäläinen. 2022. On Assessing Vulnerabilities of the 5G Networks to Adversarial Examples. *IEEE Access* 10 (2022), 126285–126303. https://doi.org/10.1109/ACCESS.2022.3225921

A APPENDIX

	Publications Selected – Reference Data
1	Patil, A. A., & Badgujar, V. S. (2018). A Comprehensive Survey on Theoretic Perspective Providing Future Directions on IoT. 2018 International Conference on Smart City and Emerging Technology (ICSCET), 1–7. https://doi.org/10.1109/ICSCET.2018.8537285
2	Yue, Y., Ming, Z., Zhijie, Q., Lei, L., & Hong, C. (2020). A Data Protection-Oriented Design Procedure for a Federated Learning Framework. 2020 International Conference on Wireless Communications and Signal Processing (WCSP), 968–974. https://doi.org/10.1109/WCSP49889.2020.9299730
3	Otoum, S., Ridhawi, I. al, & Mouftah, H. (2023). A Federated Learning and Blockchain-Enabled Sustainable Energy Trade at the Edge: A Framework for Industry 4.0. IEEE Internet of Things Journal, 10(4), 3018–3026. https://doi.org/10.1109/JIOT.2022.3140430
4	Pradhan, K. B., Sarbhadhikari, S. N., & John, P. (2021). A Framework of Responsible Innovation (RI) Model for Artificial Intelligence (AI) in Indian Healthcare. Online Journal of Health & Allied Sciences, 20(2), 1–3. https://search.ebscohost.com/login.aspx?direct=true&db=cin20&AN=152340901&site=ehost-live
5	Chukkapalli, S. S. L., Ranade, P., Mittal, S., & Joshi, A. (2021). A Privacy Preserving Anomaly Detection Framework for Cooperative Smart Farming Ecosystem. 2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 340–347. https://doi.org/10.1109/TPSISA52974.2021.00037
6	Zheng, X., Zhang, L., Hui, B., Tian, L., & Cai, Z. (2022). A Secure and Efficient Framework for Multi-Round Data Trading Over the Internet of Artificially Intelligent Things. IEEE Internet of Things Magazine, 5(1), 119–124. https://doi.org/10.1109/IOTM.001.2100194
7	Khowaja, S. A., Khuwaja, P., Dev, K., Lee, I. H., Khan, W. U., Wang, W., Qureshi, N. M. F., & Magarini, M. (2023). A Secure Data Sharing Scheme in Community Segmented Vehicular Social Networks for 6G. IEEE Transactions on Industrial Informatics, 19(1), 890–899. https://doi.org/10.1109/TII.2022.3188963
8	Li, J., Cai, J., Khan, F., Rehman, A. U., Balasubramaniam, V., Sun, J., & Venu, P. (2020). A Secured Framework for SDN-Based Edge Computing in IoT-Enabled Healthcare System. IEEE Access, 8, 135479–135490. https://doi.org/10.1109/ACCESS.2020.3011503
9	Cai, Q., Wang, H., Li, Z., & Liu, X. (2019). A Survey on Multimodal Data-Driven Smart Healthcare Systems: Approaches and Applications. IEEE Access, 7, 133583–133599. https://doi.org/10.1109/ACCESS.2019.2941419
10	Das, A. K., Bera, B., & Giri, D. (2021). AI and Blockchain-Based Cloud-Assisted Secure Vaccine Distribution and Tracking in IoMT-Enabled COVID-19 Environment. IEEE Internet of Things Magazine, 4(2), 26–32. https://doi.org/10.1109/IOTM.0001.2100016
11	Sadeghi, K., Banerjee, A., & Gupta, S. K. S. (2019). An Analytical Framework for Security-Tuning of Artificial Intelligence Applications Under Attack. 2019 IEEE International Conference On Artificial Intelligence Testing (AITest), 111–118. https://doi.org/10.1109/AITest.2019.00012
12	Zebin, T., Rezvy, S., & Luo, Y. (2022). An Explainable AI-Based Intrusion Detection System for DNS Over HTTPS (DoH) Attacks. IEEE Transactions on Information Forensics and Security, 17, 2339–2349. https://doi.org/10.1109/TIFS.2022.3183390
13	Lu, J., Xiang, X., Shen, D., Chen, G., Chen, N., Blasch, E., Pham, K., & Chen, Y. (2018). Artificial intelligence based directional mesh network design for spectrum efficiency. 2018 IEEE Aerospace Conference, 1–9. https://doi.org/10.1109/AERO.2018.8396558
14	Vuppalapati, C., Ilapakurti, A., Chillara, K., Kedari, S., & Mamidi, V. (2020). Automating Tiny ML Intelligent Sensors DevOPS Using Microsoft Azure. 2020 IEEE International Conference on Big Data (Big Data), 2375–2384. https://doi.org/10.1109/BigData50022.2020.9377755
15	Wang, R., Xu, J., Ma, Y., Talha, M., Al-Rakhami, M. S., & Ghoneim, A. (2021). Auxiliary Diagnosis of COVID-19 Based on 5G-Enabled Federated Learning. IEEE Network, 35(3), 14–20. https://doi.org/10.1109/MNET.011.2000704
16	Lin, X., Wu, J., Bashir, A. K., Li, J., Yang, W., & Piran, Md. J. (2022). Blockchain-Based Incentive Energy-Knowledge Trading in IoT: Joint Power Transfer and AI Design. IEEE Internet of Things Journal, 9(16), 14685–14698. https://doi.org/10.1109/JIOT.2020.3024246
17	Rahman, Z., Khalil, I., Yi, X., & Atiquzzaman, M. (2021). Blockchain-Based Security Framework for a Critical Industry 4.0 Cyber-Physical System. IEEE Communications Magazine, 59(5), 128–134. https://doi.org/10.1109/MCOM.001.2000679
18	Sheeraz, M. M., Athar, A., Hussain, A., Aich, S., Joo, M.-I., & Kim, H.-C. (2021). Blockchain, AI & IoT Based COVID-19 Contact Tracing and Distancing Framework. 2021 International Conference on Robotics and Automation in Industry (ICRAI), 1–6. https://doi.org/10.1109/ICRAI54018.2021.9651350
19	Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Fig. 16. Systematic Map - Selected Papers (Part 1)

20	Kang, J., Li, X., Nie, J., Liu, Y., Xu, M., Xiong, Z., Niyato, D., & Yan, Q. (2022). Communication-Efficient and Cross-Chain Empowered Federated Learning for Artificial Intelligence of Things. IEEE Transactions on Network Science and Engineering, 9(5), 2966–2977. https://doi.org/10.1109/TNSE.2022.3178970
21	Chen, J., Wang, J., Peng, T., Sun, Y., Cheng, P., Ji, S., Ma, X., Li, B., & Song, D. (2022). Copy, Right? A Testing Framework for Copyright Protection of Deep Learning Models. 2022 IEEE Symposium on Security and Privacy (SP), 824–841. https://doi.org/10.1109/SP46214.2022.9833747
22	Witt, L., Heyer, M., Toyoda, K., Samek, W., & Li, D. (2023). Decentral and Incentivized Federated Learning Frameworks: A Systematic Literature Review. IEEE Internet of Things Journal, 10(4), 3642–3663. https://doi.org/10.1109/JIOT.2022.3231363
23	Brayford, D., Vallecorsa, S., Atanasov, A., Baruffa, F., & Riviera, W. (2019). Deploying AI Frameworks on Secure HPC Systems with Containers. 2019 IEEE High Performance Extreme Computing Conference (HPEC), 1–6. https://doi.org/10.1109/HPEC.2019.8916576
24	Pavlidis, G. (2023). Deploying artificial intelligence for anti-money laundering and asset recovery: the dawn of a new era. Journal of Money Laundering Control, 26(7), 155–166. https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=163820805&site=ehost-live
25	Chen, H., Hussain, S. U., Boemer, F., Stapf, E., Sadeghi, A. R., Koushanfar, F., & Cammarota, R. (2020). Developing Privacy-preserving AI Systems: The Lessons learned. 2020 57th ACM/IEEE Design Automation Conference (DAC), 1–4. https://doi.org/10.1109/DAC18072.2020.9218662
26	Yavuz, A. A., Nouma, S. E., Hoang, T., Earl, D., & Packard, S. (2022). Distributed Cyber-infrastructures and Artificial Intelligence in Hybrid Post-Quantum Era. 2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA), 29–38. https://doi.org/10.1109/TPS-ISA56441.2022.00014
27	Xu, Y., Guo, R., Liu, X., Luo, H., Dong, C., Yao, A., & Li, X. (2022). Efficient Face Recognition via Multi-UAV-Edge Collaboration in UAV Delivery Service. 2022 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), 676–683. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom57177.2022.00092
28	Alonso-Moral, J. M., Mencar, C., & Ishibuchi, H. (2022). Explainable and Trustworthy Artificial Intelligence [Guest Editorial]. IEEE Computational Intelligence Magazine, 17(1), 14–15. https://doi.org/10.1109/MCI.2021.3129953
29	Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., & Zomaya, A. Y. (2022). Federated Learning for COVID-19 Detection With Generative Adversarial Networks in Edge Cloud Computing. IEEE Internet of Things Journal, 9(12), 10257–10271. https://doi.org/10.1109/JIOT.2021.3120998
30	Wang, Y., Su, Z., Luan, T. H., Li, R., & Zhang, K. (2022). Federated Learning With Fair Incentives and Robust Aggregation for UAV-Aided Crowdsensing. IEEE Transactions on Network Science and Engineering, 9(5), 3179–3196. https://doi.org/10.1109/TNSE.2021.3138928
31	Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. ACM Trans. Intell. Syst. Technol., 10(2). https://doi.org/10.1145/3298981
32	Huang, L., Wei, X., Zhu, P., Gao, Y., Chen, M., & Kang, B. (2020). Federated Quantile Regression over Networks. 2020 International Wireless Communications and Mobile Computing (IWCMC), 57–62. https://doi.org/10.1109/IWCMC48107.2020.9148186
33	Qin, W., Yang, L., & Ma, J. (2021). FedGR: A Lossless-Obfuscation Approach for Secure Federated Learning. 2021 IEEE Global Communications Conference (GLOBECOM), 1–6. https://doi.org/10.1109/GLOBECOM46510.2021.9686029
34	Bhagavan, S., Gharibi, M., & Rao, P. (2021). FedSmarteum: Secure Federated Matrix Factorization Using Smart Contracts for Multi-Cloud Supply Chain. 2021 IEEE International Conference on Big Data (Big Data), 4054–4063. https://doi.org/10.1109/BigData52589.2021.9671789
35	Yang, H., He, H., Zhang, W., & Cao, X. (2021). FedSteg: A Federated Transfer Learning Framework for Secure Image Steganalysis. IEEE Transactions on Network Science and Engineering, 8(2), 1084–1094. https://doi.org/10.1109/TNSE.2020.2996612
36	Roosan, D., Wu, Y., Tatla, V., Li, Y., Kugler, A., Chok, J., & Roosan, M. R. (2022). Framework to enable pharmacist access to health care data using Blockchain technology and artificial intelligence. Journal of the American Pharmacists Association: JAPhA, 62(4), 1124–1132. https://search.ebscohost.com/login.aspx?direct=true&db=cin20&AN=157819570&site=ehost-live
37	Wazid, M., Das, A. K., Mohd, N., & Park, Y. (2022). Healthcare 5.0 Security Framework: Applications, Issues and Future Research Directions. IEEE Access, 10, 129429–129442. https://doi.org/10.1109/ACCESS.2022.3228505
38	Su, L., & Lau, V. K. N. (2021). Hierarchical Federated Learning for Hybrid Data Partitioning Across Multitype Sensors. IEEE Internet of Things Journal, 8(13), 10922–10939. https://doi.org/10.1109/JIOT.2021.3051382

Fig. 17. Systematic Map - Selected Papers (Part 2)

39	Zhang, Z., Huang, L., Tang, R., Peng, T., Guo, L., & Xiang, X. (2020). Industrial Blockchain of Things: A Solution for Trustless Industrial Data Sharing and Beyond. 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), 1187–1192. https://doi.org/10.1109/CASE48305.2020.9216817
40	Liu, C., Zhang, L., Dai, Y., Chen, F., Chen, H., & Zhong, P. (2022). Intel SGX-Based Trust Framework Designed for Secure Machine Learning. 2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2), 1621–1627. https://doi.org/10.1109/EI256261.2022.10116101
41	Babbar, H., Rani, S., & AlQahtani, S. A. (2022). Intelligent Edge Load Migration in SDN-IIoT for Smart Healthcare. IEEE Transactions on Industrial Informatics, 18(11), 8058–8064. https://doi.org/10.1109/TII.2022.3172489
42	Parfenov, D., Grishina, L., Legashev, L., Zhigalov, A., & Parfenov, A. (2023). Investigation of the Security of ML-models in IoT Networks from Adversarial Attacks. 2023 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), 229–232. https://doi.org/10.1109/USBEREIT58508.2023.10158812
43	Li, H., Meng, D., Wang, H., & Li, X. (2020). Knowledge Federation: A Unified and Hierarchical Privacy-Preserving AI Framework. 2020 IEEE International Conference on Knowledge Graph (ICKG), 84–91. https://doi.org/10.1109/ICBK50248.2020.00022
44	Wang, Y., Su, Z., Zhang, N., & Benslimane, A. (2021). Learning in the Air: Secure Federated Learning for UAV-Assisted Crowdsensing. IEEE Transactions on Network Science and Engineering, 8(2), 1055–1069. https://doi.org/10.1109/TNSE.2020.3014385
45	Zolotukhin, M., Miraghaei, P., Zhang, D., & Hämäläinen, T. (2022). On Assessing Vulnerabilities of the 5G Networks to Adversarial Examples. IEEE Access, 10, 126285–126303. https://doi.org/10.1109/ACCESS.2022.3225921
46	Pandl, K. D., Thiebes, S., Schmidt-Kraepelin, M., & Sunyaev, A. (2020). On the Convergence of Artificial Intelligence and Distributed Ledger Technology: A Scoping Review and Future Research Agenda. IEEE Access, 8, 57075–57095. https://doi.org/10.1109/ACCESS.2020.2981447
47	Wazid, M., Das, A. K., Shetty, S., & Rodrigues, J. J. P. C. (2020). On the Design of Secure Communication Framework for Blockchain-Based Internet of Intelligent Battlefield Things Environment. IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 888–893. https://doi.org/10.1109/INFOCOMWKSHPS50562.2020.9163066
48	Zhang, X., Wang, Y., Lu, S., Liu, L., xu, L., & Shi, W. (2019). OpenEI: An Open Framework for Edge Intelligence. 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), 1840–1851. https://doi.org/10.1109/ICDCS.2019.00182
49	Alberti, A. M., Santos, M. A. S., Souza, R., da Silva, H. D. L., Carneiro, J. R., Figueiredo, V. A. C., & Rodrigues, J. J. P. C. (2019). Platforms for Smart Environments and Future Internet Design: A Survey. IEEE Access, 7, 165748–165778. https://doi.org/10.1109/ACCESS.2019.2950656
50	Shao, Y., Tian, C., Han, L., Xian, H., & Yu, J. (2022). Privacy-Preserving and Verifiable Cloud-Aided Disease Diagnosis and Prediction With Hyperplane Decision-Based Classifier. IEEE Internet of Things Journal, 9(21), 21648–21661. https://doi.org/10.1109/JIOT.2022.3181734
51	Fritchman, K., Saminathan, K., Dowsley, R., Hughes, T., de Cock, M., Nascimento, A., & Teredesai, A. (2018). Privacy-Preserving Scoring of Tree Ensembles: A Novel Framework for AI in Healthcare. 2018 IEEE International Conference on Big Data (Big Data), 2413–2422. https://doi.org/10.1109/BigData.2018.8622627
52	Otoum, S., Ridhawi, I. al, & Mouftah, H. (2022). Realizing Health 4.0 in Beyond 5G Networks. ICC 2022 - IEEE International Conference on Communications, 2960–2965. https://doi.org/10.1109/ICC45855.2022.9838687
53	Confido, A., Ntagiou, E. v, & Wallum, M. (2022). Reinforcing Penetration Testing Using AI. 2022 IEEE Aerospace Conference (AERO), 1–15. https://doi.org/10.1109/AERO53065.2022.9843459
54	Cheng, Y., Meng, H., Lei, Y., & Tan, X. (2021). Research on Privacy Protection Technology in Face Identity Authentication System Based on Edge Computing. 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID), 438–449. https://doi.org/10.1109/AIID51893.2021.9456477
55	Das, P., Illa, M., Pokhariyal, R., Latoria, A., Hemlata, & Saini, D. J. B. (2023). Role of Neural Network, Fuzzy, and IoT in Integrating Artificial Intelligence as a Cyber Security System. 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), 652–658. https://doi.org/10.1109/ICEARS56392.2023.10084988
56	Uddin, R., & Kumar, S. (2022). SDN-based Federated Learning approach for Satellite-IoT Framework to Enhance Data Security and Privacy in Space Communication. 2022 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE), 71–76. https://doi.org/10.1109/WiSEE49342.2022.9926943
57	Su, Z., Wang, Y., Luan, T. H., Zhang, N., Li, F., Chen, T., & Cao, H. (2022). Secure and Efficient Federated Learning for Smart Grid With Edge-Cloud Collaboration. IEEE Transactions on Industrial Informatics, 18(2), 1333–1344. https://doi.org/10.1109/TII.2021.3095506
58	Xu, M., Hoang, D. T., Kang, J., Niyato, D., Yan, Q., & Kim, D. I. (2022). Secure and Reliable Transfer Learning Framework for 6G-Enabled Internet of Vehicles. *IEEE Wireless Communications*, *29*(4), 132–139. https://doi.org/10.1109/MWC.004.2100542

Fig. 18. Systematic Map - Selected Papers (Part 3)

59	Tang, X., Zhu, L., Shen, M., Peng, J., Kang, J., Niyato, D., & El-Latif, A. A. A. (2022). Secure and Trusted Collaborative Learning Based on Blockchain for Artificial Intelligence of Things. IEEE Wireless Communications, 29(3), 14–22. https://doi.org/10.1109/MWC.003.2100598
60	Chakrabarty, S., & Engels, D. W. (2020). Secure Smart Cities Framework Using IoT and AI. 2020 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT), 1–6. https://doi.org/10.1109/GCAIoT51063.2020.9345912
61	Zhou, D., Yu, Y., Wu, D., Gan, Q., Chen, Z., & Xu, B. (2022). SecureAstraea: A Self-balancing Privacy-preserving Federated Learning Framework. 2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 1–8. https://doi.org/10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927969
62	Bera, B., Wazid, M., Das, A. K., & Rodrigues, J. J. P. C. (2021). Securing Internet of Drones Networks Using AI-Envisioned Smart-Contract-Based Blockchain. IEEE Internet of Things Magazine, 4(4), 68–73. https://doi.org/10.1109/IOTM.001.2100044
63	Ikharo, B., Obiagwu, A., Obasi, C., Hussein, S. U., & Akah, P. (2021). Security for Internet-of-Things Enabled E-Health using Blockchain and Artificial Intelligence: A Novel Integration Framework. 2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), 1–4. https://doi.org/10.1109/ICMEAS52683.2021.9692368
64	Mahendra, I., Ramadhan, A., Trisetyarso, A., Abdurachman, E., & Zarlis, M. (2022). Strategic Information System Planning in the Industry 4.0 Era: A Systematic Literature Review. 2022 IEEE Creative Communication and Innovative Technology (ICCIT), 1–7. https://doi.org/10.1109/ICCIT55355.2022.10119002
65	Wehrmeister, K. A., Bothos, E., Marinakis, V., Magoutas, B., Pastor, A., Carreras, L., & Monti, A. (2022). The BD4NRG Reference Architecture for Big Data Driven Energy Applications. 2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA), 1–8. https://doi.org/10.1109/IISA56318.2022.9904424
66	Minssen, T., Seitz, C., Aboy, M., & Corrales Compagnucci, M. (2020). The EU-US Privacy Shield Regime for Cross-Border Transfers of Personal Data under the GDPR: What are the legal challenges and how might these affect cloud-based technologies, big data, and AI in the medical sector?. European Pharmaceutical Law Review, 4(1), 34–50. https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=142701509&site=ehost-live
67	Suryavanshi, A., G, A., N, M. B. T., M, R., & N, A. H. (2023). The integration of Blockchain and AI for Web 3.0: A security Perspective. 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), 1–8. https://doi.org/10.1109/ICITIIT57246.2023.10068672
68	Bruce, P. C., & Fleming, G. (2021). The Responsible Data Science Framework. In Responsible Data Science (pp. 73–97). Wiley. https://ieeexplore.ieee.org/document/9942267
69	Naderi, E., & Asrari, A. (2022). Toward Detecting Cyberattacks Targeting Modern Power Grids: A Deep Learning Framework. 2022 IEEE World AI IoT Congress (AIIoT), 357–363. https://doi.org/10.1109/AIIoT54504.2022.9817309
70	Khowaja, S. A., Dev, K., Qureshi, N. M. F., Khuwaja, P., & Foschini, L. (2022). Toward Industrial Private AI: A Two-Tier Framework for Data and Model Security. IEEE Wireless Communications, 29(2), 76–83. https://doi.org/10.1109/MWC.001.2100479
71	Pinyoanuntapong, P., Huff, W. H., Lee, M., Chen, C., & Wang, P. (2022). Toward Scalable and Robust AIoT via Decentralized Federated Learning. IEEE Internet of Things Magazine, 5(1), 30–35. https://doi.org/10.1109/IOTM.006.2100216
72	Shafique, M., Marchisio, A., Wicaksana Putra, R. V., & Hanif, M. A. (2021). Towards Energy-Efficient and Secure Edge AI: A Cross-Layer Framework ICCAD Special Session Paper. 2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD), 1–9. https://doi.org/10.1109/ICCAD51958.2021.9643539
73	ARBAOUI, M., BRAHMIA, M.-E.-A., & RAHMOUN, A. (2022). Towards secure and reliable aggregation for Federated Learning protocols in healthcare applications. 2022 Ninth International Conference on Software Defined Systems (SDS), 1–3. https://doi.org/10.1109/SDS57574.2022.10062923
74	Solomon, A., & Crawford, Z. (2021). Transitioning from Legacy Air Traffic Management to Airspace Management through Secure, Cloud-Native Automation Solutions. 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), 1–8. https://doi.org/10.1109/DASC52595.2021.9594313
75	Peng, Z., Xu, J., Chu, X., Gao, S., Yao, Y., Gu, R., & Tang, Y. (2022). VFChain: Enabling Verifiable and Auditable Federated Learning via Blockchain Systems. IEEE Transactions on Network Science and Engineering, 9(1), 173–186. https://doi.org/10.1109/TNSE.2021.3050781

Fig. 19. Systematic Map - Selected Papers (Part 4)

	Lifecycle														Value-Based							Methodology		
	Inception	Data Collection	Data Cleaning	Data Processing	Model Design	Model Training	Model Testing	Optimization	Model Evaluation	Deployment	Monitoring	Re-Training	Retirement	Complete Lifecycle	Robustness	Safety	Bias	Technical Security	Alignment	Explainability	Privacy	Publication Format	Contribution Type	Evaluation Method
1	X																	X				C	P	Ep
2						X									X						X	C	S	Ep
3																						J	S	Ep
4		X	X	X											X					X	X	J	S	Ex
5		X	X	X			X														X	C	S	Ex
6						X			X									X			X	J	S	Ep
7																					X	J	S	Ep
8		X	X	X		X															X	J	S	Ep
9	X	X			X													X				J	E	LR
10		X				X									X					X	X	J	S	Ep
11						X	X	X	X						X							C	S	Ep
12											X				X			X		X		J	S	Ep
13						X												X				C	S	Ex
14														X	X							C	S	CS
15							X		X	X		X									X	J	S	Ep
16						X				X	X				X			X			X	J	S	Ep
17	X	X	X	X						X					X			X			X	J	S	Ep
18		X		X		X					X				X			X			X	C	S	Ex
19														X		X	X		X			C	S	Ep
20						X												X			X	J	S	Ep
21					X	X			X				X		X			X				C	S	Ep
22																		X			X	J	E	LR
23										X								X				C	EP	CS
24															X			X				J	E	LR
25						X	X		X		X				X			X			X	C	E	Ex
26																					X	C	S	Ep
27						X			X						X						X	C	S	Ep
28																				X	X	J	E	Ex
29	X				X	X									X			X			X	J	S	Ep
30					X	X									X						X	J	S	Ep
31	X	X				X												X			X	J	S	Ex
32	X					X					X							X			X	C	S	Ep
33					X													X				C	S	Ep
34					X										X			X			X	C	S	Ep
35						X				X											X	J	S	Ep
36														X				X			X	J	S	LR
37					X				X									X				J	SE	LR
38						X				X	X				X						X	J	S	Ep
39	X	X				X				X					X			X			X	C	S	Ex
40																		X			X	C	S	Ep

Fig. 20. Initial Systematic Map - Not Summarized (Part 1)

	Lifecycle														Value-Based							Methodology		
	Inception	Data Collection	Data Cleaning	Data Processing	Model Design	Model Training	Model Testing	Optimization	Model Evaluation	Deployment	Monitoring	Re-Training	Retirement	Complete Lifecycle	Robustness	Safety	Bias	Technical Security	Alignment	Explainability	Privacy	Publication Format	Contribution Type	Evaluation Method
41										X					X			X			X	J	S	Ep
42						X																C	S	Ep
43						X					X				X			X			X	C	S	Ep
44		X				X									X			X			X	J	S	Ep
45							X		X						X			X				J	E	Ep
46	X	X				X									X	X				X	X	J	E	LR
47															X	X		X				C	S	Ex
48	X	X			X					X					X			X			X	C	S	Ex
49										X				X				X				J	E	M
50														X	X			X			X	J	S	Ep
51										X											X	C	S	Ep
52						X	X		X									X			X	C	S	Ep
53											X				X	X		X				C	S	Ep
54					X	X			X	X	X				X			X			X	C	S	Ep
55					X	X			X						X			X				C	S	Ep
56		X				X									X			X			X	C	S	Ex
57						X															X	J	S	Ep
58						X				X		X			X	X		X				J	S	Ep
59																					X	J	S	Ep
60										X					X	X		X				C	S	Ex
61				X		X															X	C	S	Ep
62		X		X											X	X		X				J	S	Ep
63	X				X					X					X			X			X	C	S	Ex
64														X	X	X			X			C	S	LR
65		X																X	X		X	C	S	Ex
66																		X			X	J	E	CS
67																		X			X	C	S	CS
68														X		X	X		X	X		B	S	Ex
69					X					X	X				X	X		X				C	S	Ep
70		X																			X	J	S	Ep
71						X									X			X			X	J	S	Ep
72		X				X			X	X	X				X			X			X	C	S	Ex
73																		X			X	C	S	Ep
74	X				X					X					X	X		X	X			C	S	Ep
75	X					X	X		X	X					X			X			X	J	S	Ep

Fig. 21. Initial Systematic Map - Not Summarized (Part 2)