IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 June 2023, accepted 13 June 2023, date of publication 19 June 2023, date of current version 23 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287195

SURVEY

A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle

SAKIB SHAHRIAR[1], SONAL ALLANA[1], SEYED MEHDI HAZRATIFARD[2], AND ROZITA DARA[1]
[1]School of Computer Science, University of Guelph, Guelph, ON N1G 2W1, Canada
[2]Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 2Y2, Canada

Corresponding author: Rozita Dara (drozita@uoguelph.ca)

The work of Rozita Dara was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant.

ABSTRACT Over the decades, Artificial Intelligence (AI) and machine learning has become a transformative solution in many sectors, services, and technology platforms in a wide range of applications, such as in smart healthcare, financial, political, and surveillance systems. In such applications, a large amount of data is generated about diverse aspects of our life. Although utilizing AI in real-world applications provides numerous opportunities for societies and industries, it raises concerns regarding data privacy. Data used in an AI system are cleaned, integrated, and processed throughout the AI life cycle. Each of these stages can introduce unique threats to individual's privacy and have an impact on ethical processing and protection of data. In this paper, we examine privacy risks in different phases of the AI life cycle and review the existing privacy-enhancing solutions. We introduce four different categories of privacy risk, including (i) risk of identification, (ii) risk of making an inaccurate decision, (iii) risk of non-transparency in AI systems, and (iv) risk of non-compliance with privacy regulations and best practices. We then examined the potential privacy risks in each AI life cycle phase, evaluated concerns, and reviewed privacy-enhancing technologies, requirements, and process solutions to countermeasure these risks. We also reviewed some of the existing privacy protection policies and the need for compliance with available privacy regulations in AI-based systems. The main contribution of this survey is examining privacy challenges and solutions, including technology, process, and privacy legislation in the entire AI life cycle. In each phase of the AI life cycle, open challenges have been identified.

INDEX TERMS Artificial intelligence, machine learning, AI life cycle, privacy risk, privacy legislation, privacy enhancing solutions.

I. INTRODUCTION

Artificial intelligence (AI) refers to the development of computational agents that can perform tasks associated with human intelligence, including speech recognition, visual perception, and general problem solving. An AI system is expected to be able to attain human performance and be rational by doing the "right thing" given the available information [1]. Turing was one of the first to challenge the ability of computer systems to match human reasoning by asking "Can machines think?" [2] and consequently developing the Turing Test. If a human interrogator is incapable of distinguishing between the answers of a person and a computer to a series of written questions, then the computer system passes the Turing Test. Machine learning (ML) is a branch of AI where computer systems learn from experience, i.e., from the given data, without explicit programming. Upon successful learning, these robust models can offer intelligent decisions and improve different dimensions of our daily lives. For instance, in the healthcare industry, image recognition and analysis by ML in applications such as cancer diagnosis [3] can help physicians make diagnostic decisions. Other examples include the use of AI in applications such as forecasting stock price variations [4] as well as predicting the progression and vaccine development for contagious diseases such as COVID-19 [5]. To successfully implement AI in

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleylek.

these applications, AI algorithms require computing power, efficient algorithms, and most importantly representative data to learn from.

Since AI is fundamentally a data-driven approach, allowing AI systems to access and process our private information in many day-to-day applications is inevitable. For instance, people have to submit their personal and financial information to determine their eligibility for financial support, such as mortgages and business loans. Another example is the collection of cookies and browsing history to provide personalized advertising when visiting a website [6]. This increasing trend in accessing and processing personal data has raised concerns about data privacy. The reason for privacy concerns is due to the fact that data related to people can be sensitive. Even if they do not contain explicit and direct information about their identity, AI methods can be used to extract sensitive personal information from individuals' data. A common approach towards solving this problem is to anonymize records containing sensitive information. For instance, a dataset containing 100 million anonymous movie ratings was released by Netflix in 2006 as part of an open contest to develop an accurate recommendation system. However, this dataset was de-anonymized in 2008 by researchers utilizing only a small amount of information from other public databases [7]. The researchers also concluded that the revealed information could potentially identify sensitive information about users, including their political and religious beliefs.

Loss of privacy can have a devastating impact on individuals. Loss of reputation, identity theft, biased and unfair decisions, and legal consequences are only some of the consequences of privacy breaches and concerns. To guarantee effective privacy preservation mechanisms, organizations need to consider privacy as an integral component in developing their technologies and managing sensitive data. This principle is also known as privacy by design, where the objective is to proactively integrate solutions to prevent privacy in the development, operation, and management phases of information processing technologies [8]. In contrast, privacy by policy provides guidelines for preserving privacy and includes mechanisms such as informing users about how their data is used and enforcing compliance with privacy legislation. Incorporating privacy into the architecture of technology, i.e., by design, is considered more reliable than applying privacy by policy [9]. There is also an increasing trend in establishing regulations that impact the design and development of technology solutions such as AI systems. For instance, the interpretability of AI algorithms and processes has been the center of attention by some regulations, such as the Information Commissioner's Office (ICO) in the United Kingdom. A comprehensive understanding of the various regulatory and technical requirements in different developmental stages is required in order to apply privacy by design to AI systems. This is because privacy can be compromised in every phase of the AI development lifecycle and privacy-preserving solutions must be incorporated at every stage. Additionally, to develop new and effective privacy-preserving solutions, it is essential to investigate existing solutions and identify their limitations. Finally, the emergence of large language models like ChatGPT present significant privacy challenges due to their ability to access and potentially reveal personal data. To mitigate these challenges, robust data privacy and security policies must be developed and implemented to ensure compliance with regulations such as GDPR [10]. Regular audits of these policies are necessary to identify and address any potential vulnerabilities. Additionally, models must be trained on data that does not contain personal information, and measures must be taken to prevent unintentional leakage of personal data.

A comprehensive analysis of the existing literature is required to understand the research gaps in privacy-focused AI development. Several surveys have made attempts to review privacy threats in intelligent systems. Liu et al. [11] investigated privacy preservation challenges in ML, focusing on deep learning, algorithms that utilize deep neural networks. Although the paper provides a broad review of different privacy attacks, it does not consider important phases of AI and ML development, including planning and data collection. Moreover, [12] discusses various cyber security threats with an emphasis on adversarial learning and several attacks and defense mechanisms across ML algorithms. Despite the comprehensive presentation of security threats, the authors did not address privacy concerns from an AI lifecycle perspective. Boulemtafes et al. [13] reviewed state-of-the-art solutions that deal with preserving the privacy of deep learning algorithms. In the learning phase, leakage of training data and model parameters were identified as threats to privacy. In the model analysis and deployment phases, the concerns were with the release of sensitive information and model parameters, respectively. Similarly, [14] investigated several privacy and security concerns of deep learning models. In terms of privacy, the focus was on model extraction and model inversion attacks. However, both [13] and [14] do not address the impact of privacy policies on AI development and also fail to tackle some of the important phases in the AI lifecycle. In another survey, Ashmore et al. [15] investigated the approaches to ensure the ML algorithms are safe for deployment. By considering the various ML lifecycle, the authors highlighted the important steps to provide assurance, i.e., ensure the ML models are accurate for their intended purpose. However, the privacy aspects in the context of the AI life cycle were beyond the scope of this work. Wickramasinghe et al. [16] analyzed the interactions between AI systems, developers, and users in the AI Life Cycle to enhance the trustworthiness of AI. Although some aspects of privacy were discussed, the focus of this work was on highlighting important principles to facilitate trust in AI.

Although various existing works have reviewed privacy and security challenges related to AI and ML systems, there is a necessity for examining and addressing privacy concerns

in AI systems and during all phases of the AI life cycle. In addition to focusing only on certain phases of the AI life cycle, existing surveys also pay little attention to privacy preservation strategies. Consequently, to bridge the gap, this survey provides an overview of the AI life cycle phases and investigates privacy risks along with their corresponding mitigation strategies. Following are the main contributions of this paper:

- Proposes a novel privacy risk categorization framework consisting of four groups: risk of identification, risk of making inaccurate decisions, risk of non-transparent AI systems, and risk of non-compliance with privacy regulations.
- Discusses privacy solutions addressing the aforementioned privacy risks at each stage of AI lifecycle.
- Addresses the impact of data protection practices and legislation requirements to address privacy risks in the AI life cycle phases.
- Discusses future research directions and open challenges in the context of privacy in the AI lifecycle.

The focus of this paper is on the AI lifecycle, encompassing various branches of AI, including ML, expert systems, and natural language processing (NLP). Expert systems are computer programs that mimic the decision-making ability of human experts in a specific domain, while NLP involves the use of algorithms and models to analyze and interpret text or speech data. By examining these branches of AI through the lens of the AI lifecycle, this paper provides a comprehensive understanding of the potential applications and challenges of deploying AI systems in different domains.

The rest of this survey is organized as follows. Section II investigates the need for privacy in AI systems. Section III presents the AI life cycle concisely and Section IV categorizes privacy risks. Section V highlights privacy risks in each AI life cycle phase and their solutions. Section VI discusses the impact of regulations on AI development and hints at future research directions. Finally, Section VI concludes the paper.

II. THE NEED FOR PRIVACY IN ARTIFICIAL INTELLIGENCE SYSTEMS

Privacy is an essential legal and social concept that gives people control over who has access to their sensitive information. Margulis [17] discussed the work of two theorists, Alan Westin and Irwin Altman, and their impacts on the theory of privacy. According to Margulis, Westin's theory of privacy revolves around temporarily restricting access to individuals to protect themselves whereas Altman's theory of privacy deals with selectively controlling access to individuals to preserve privacy. Margulis also argues that based on the two theories, privacy can be understood as a psychological concept. Although these standard definitions provide a foundation for developing privacy frameworks, there is a need to evolve into a modern approach for privacy, i.e., information privacy [18], that is more suitable for technological development. Information privacy can be defined as ''the ability of the individual to control information about one's self'' [19]. According to International Organization for Standardization (ISO), confidentiality (an attribute of privacy) is defined as ''the property, that information is not made available or disclosed to unauthorized individuals, entities, or processes'' [20]. The two examples of privacy definitions indicate a lack of a standard definition of privacy that explains the privacy requirements and risks in information systems. The complex and diverse nature of privacy risks in developing AI systems makes it even more challenging in addressing them. Therefore, a suitable framework for categorizing privacy risks in light of AI development is needed. Organizing privacy risks into appropriate categories will help developers understand them better and make risk-informed decisions [21]. Privacy risk categorization will also help developers integrate privacy-preserving solutions into AI development.

The rapid development in technology, including AI, is complicating attempts to safeguard privacy in such systems. AI systems are inherently data-driven, and considering their potential to extract hidden information in data, protecting privacy in the context of AI requires sophisticated approaches [22]. AI algorithms are generally trained using high dimensional data. This high dimensional data can contain numerous attributes of an individual, increasing the risk of identifying an individual by cross-referencing with other public datasets [23]. This is a threat to privacy because, in most cases, individuals agreed to share their personal data for aggregate analysis (in this case, to train the AI system). However, the identified attributes of an individual may be used for malicious purposes both online and offline, including fraud and harassment. Moreover, AI algorithms can be utilized by attackers to infer sensitive information about data subjects, such as their gender and political views [24]. The wide adoption and deployment of AI models online also poses a threat to privacy. For instance, it is possible to determine whether a person has a disease by looking at or analyzing the clinical records of that person used to train an AI algorithm to model that disease [25]. Given that the threat to privacy in AI models can emerge at different phases of development, including data collection and post-deployment, a comprehensive approach is required in developing privacy-focused AI systems. According to a resolution passed by the 32nd International Conference of Data Protection and Privacy Commissioners, to completely protect privacy in any system, privacy must be embedded into the design, operation, and management of the system across its entire life cycle [26]. Consequently, to develop privacy-focused AI systems, it is necessary to analyze and resolve the privacy threats associated with each stage of the AI lifecycle. Besides developing privacy-preserving solutions into the technology, developers should also implement the guidelines defined by various legislation and policies.

Many privacy legislation and best practices are dedicated to data collection and processing requirements to protect

individuals' privacy. The European Union (EU) introduced the General Data Protection Regulation (GDPR) in 2016, which defines personal data as any information about an identified or identifiable individual (Article 4) [27]. This means any information that can potentially identify an individual is considered personal data and should be protected. Organization for Economic Co-operation and Development (OECD) [28], California Consumer Privacy Act (CCPA) [29], and Information Protection and Electronic Documents Act (PIPEDA) [30] have similar definitions for personal data. According to most privacy regulations, including GDPR and PIPEDA, personal data is also referred to as Personally Identifiable Information (PII) [31]. PII includes information related to an identified or identifiable person, therefore signifying that any information linked to an individual is subject to the privacy protection regulations [32]. Examples of sensitive PII that can directly identify an individual include social security or insurance numbers, driver's licenses, and biometric records [33]. On the other hand, information such as age and gender may be more accessible but cannot identify a person on their own [34]. However, a study has shown that 87% of the U.S. population (using the 1990 census) can be uniquely identified by combining their 5-digit ZIP, gender, and birth date [35]. In addition to these two categories, there are other data types that can identify an individual if they show unique patterns and are linked to a person. For example, mobility data of individuals may be used to infer their address, people or places they interact with, and their leisure activities, while the accuracy of such inferences continues to grow with data availability [36]. Moreover, in the above legislation and best practices, an individual has some rights to control when or how their data can be collected, used, or shared. For instance, OECD states that personal data should be protected against risks such as unauthorized access, use, modification, destruction, or disclosure. The introduction of different data privacy legislation across the globe [37] further signifies the need to secure privacy in emerging technologies such as AI.

The introduction of privacy regulations has significant impacts on developing AI systems. The regulations set by GDPR extend to the processing of personal data of any individual in the EU, implying that the activities of many foreign companies fall into the scope of GDPR [38]. These cross-border privacy rules impose a significant barrier when it comes to data sharing stages of the AI life cycle will be presented in the next section.

III. AI DEVELOPMENT LIFE CYCLE

The objective of an AI system is to solve sophisticated problems in a data-driven approach and without the need for explicit human programming. For instance, an AI system can be used to provide diagnosis from medical scans almost instantly and therefore help radiologists decrease their diagnosis time. AI systems learn from data to make decisions or predictions in specific applications. While historical data is often used to train ML algorithms, other types of AI systems may incorporate real-time data or human input to make decisions. The data are collected, preprocessed, analyzed, and utilized in different stages of AI lifecycle. In this context, data subjects are individuals whose information has been collected for the development of AI systems. The GDPR also defines a data controller as someone who determines the purpose and means of personal data processing [39]. The data controller may also authorize a data processor to possess and process personal data on behalf of the controller.

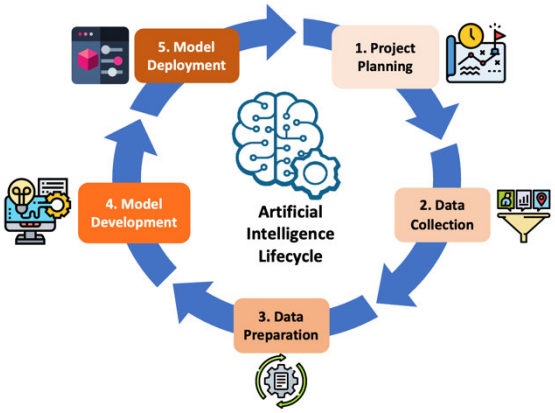

FIGURE 1. Artificial Intelligence Life Cycle.

The development of an AI system can be broken down into the AI development life cycle, which refers to the cyclical process that defines the steps to build and use an AI system [15]. We refer to this process as the AI life cycle. Figure 1 shows the schematic representation of the AI life cycle in five phases. Each of these five phases is discussed next.

A. PROJECT PLANNING

Prior to any model development, it is necessary to outline the objectives of the AI system. This includes defining the scope of the system and identifying the relevant use cases that the AI will address. Therefore, understanding the project objectives and deciding on the required data is the fundamental step of an AI life cycle. To develop an efficient AI application, it is necessary to gather comprehensive information about the project objectives and other development details, such as data sources and potential system users [40]. Failure to identify the project objectives will lead to inevitable delays in implementation and potentially impact the performance negatively. In the case of medical image diagnosis, for instance, a broader scope of identifying different kinds of injuries from magnetic resonance imaging (MRI) scans, including brain and spinal cord injuries may not be efficient. Instead, based on the data availability, training the AI model to identify a specific injury may lead to more accurate results.

The planning phase also requires identifying different skill sets to support and implement each step of the AI life cycle [41]. Employing diverse experts such as privacy professionals, ethicists, testers, AI developers, data scientists, and subject-matter experts can facilitate the development

and review of the system and process requirements. These requirements may include complying with information privacy and other legislation to ensure no violations are taking place as a result of the project's implementation. Moreover, the planning phase also ensures the quality of the AI system meets the necessary requirement. This is achieved by identifying comprehensive testing for the model as well as identifying metrics and benchmarks to compare the performance of the model upon training.

B. DATA COLLECTION

Once the objectives have been identified and the scope of the project is defined, the next phase of the AI life cycle is data collection. Since AI and ML systems are inherently data-driven, the acquisition of quality data remains an important phase. This phase provides the information needed by data scientists to implement the use cases and build the AI model [41]. To enhance the generalization of the AI models and avoid bias in the deployed model [42], collected data should cover a diverse representation of the intended statistical and problem domain. Furthermore, collected data may be in several types, such as numerical, categorical, time series, and text [43]. Each data type has its specifications and complexities for storage, processing, and maintenance that may impact the risk of vulnerability. To address this issue, an AI system should follow a standard for securing data storage from different sources. In addition to the required data, further information about the data, referred to as metadata, is collected in catalogs to facilitate the organization and usage of the data. Metadata can also hold information about access rights, data ownership, data controllers, third parties, usage purpose, retention, or other relevant information about the maintenance of data and the AI system relevant to privacy concerns [44]. Therefore, this phase should address the requirements for both data and metadata.

The data collection phase can also be time consuming and costly for the project. Identifying existing data sources can often speed up the data collection step. However, projects may require specific and new data to implement the necessary use cases. Therefore, the data collection phase must safeguard the quality of the collected data in terms of its accuracy and relevance. Moreover, given that the quality of the collected data can impact model performance [45], greater attention is required at this stage to meet the quality requirements. Furthermore, ethical issues related to data collection must also be addressed. For instance, data collection in the context of human or animal experiments is subject to ethical approval. Therefore, in addition to the quality of the data, this phase must ensure that the collected data does not violate privacy and ethical regulations.

C. DATA PREPARATION

The steps taken in preparing the data can impact both the performance and training time of the model. Depending on data types, objectives of a project, and model requirements, preprocessing steps such as feature selection, feature extraction, data integration, and data cleaning can enhance the system performance [40]. Feature selection and extraction can help deal with high-dimensional data and avoid over fitting in model training [46]. Moreover, in this phase, data scientists decide how to deal with incomplete data, missing values, outliers, and anomalous instances. For example, visualization of data characteristics can help data scientists to detect and remove anomalous samples [43] and consequently improve model performance.

Data integration is a system requirement that provides users with a standard data format residing in different sources by combining and transforming data into a single coherent store in heterogeneous conditions. This requires semantic interoperability, which refers to the ability of systems to exchange and use data in a uniform platform and make it possible to work with data from different sources [47]. For example, when two similar companies need to merge their databases, the available metadata makes it possible to reach a uniform structure for data integration [48]. Furthermore, annotating the collected samples can facilitate tasks, such as classification and association rule mining. To ensure the annotation process is accurate and reliable, it is recommended to use independent annotators. The annotators should ideally be experts in the domain, and if this is not possible, the annotation must be validated by experts. Also, data annotation should be free of discrimination to avoid building a biased model [42]. In cases of dealing with imbalanced datasets, effective strategies, including oversampling and decomposition methods [49], should be considered.

D. MODEL DEVELOPMENT

Suitable model selection and development based on the system requirements are necessary to obtain strong performance. Some algorithms, such as k-nearest neighbors, rely on the extracted rules from samples and do not require a training or learning phase. Conversely, ML-based models learn from data in a training phase by optimizing an objective function. As depicted in Figure 2, most ML algorithms include three components: input, output, and a model [43]. In supervised learning, the inputs of the model are the extracted features, and the outputs of the model are the labels or predictions. If the output of the model is a specific category, the model is a classification model and conversely, if the output is a continuous value, the model is a regression model. The model's parameters is learned during training to extract the relationships between input and output. In cases where annotated data is not available, unsupervised learning such as clustering, dimensionality reduction, and anomaly detection can be utilized. These methods are designed to find unknown structures of the input samples. In this context, the type of learning and model selection will depend on the dataset available and the problem being addressed.

The model's performance depends to a large extent on the available data and the training procedure. Training a model

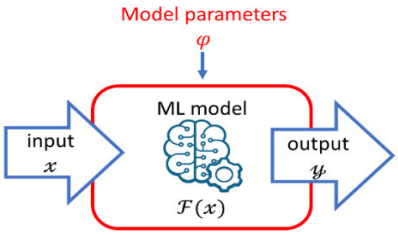

FIGURE 2. Simple Schema of an ML Model.

with limited data and complexity leads to a simple model that is not capable of accurate predictions, i.e., under fitting. On the other hand, excessive training may lead to over fitting, reducing the model's accuracy on unseen data, i.e., poor generalization. Over fitting can also make the model vulnerable to various forms of attacks [50]. In addition to model performance, transparency and interpretability can increase data subjects' trust in the models [51]. Interpretable AI refers to algorithms with understandable behavior [52], which allows data scientists to better comprehend why specific results have been obtained.

An important subphase of model development is model evaluation. This stage allows developers to adjust model parameters to maximize performance gain. A subset of the training set, called validation set, is used to assess the performance during the training. Developers may also analyze the impact of specific input features on model performance and consequently augment or deduct input features. The selection of appropriate evaluation metrics depend on the type of model and problem. For instance, there are well-defined evaluation metrics for classification and regression problems [53]. Clustering algorithms have well-defined metrics [54] based on whether some ground truth (annotated data) is available. Finally, it is important to assess the performance of the model on unseen or test data. This step highlights whether the ML model is capable of making generalized predictions.

E. MODEL DEPLOYMENT

The satisfaction of stakeholders and data subjects is highly related to the outcome of the deployed product. Therefore, the final system should be aligned with the objectives that have been planned in the first phase. Performing excessive tests before deployment can help assure the system's accuracy and ensure the model is free of bias. Successful deployment should also consider ease of use for end users. For instance, a suitable application can be developed that allows users to interact with the final model intuitively. In some cases, such as detecting traffic violations, real-time use of the AI model may be required, and such requirements must also be met at this stage. Model deployment should also ensure the necessary computing and memory resources are available for the AI algorithm to function. For example, complex models may not be suitable to be deployed on limited hardware devices, including mobile phones, and are better suited to be deployed on the cloud.

After deployment, data distribution may change, end-user preferences can evolve, and end-user feedback will emerge, highlighting the importance of regular system maintenance. While concept drift and changes in the distribution of input data or the target variable can impact the performance of the deployed model [55], it is necessary to keep a product up to date. Since many users have access to the deployed model, this phase is particularly vulnerable to attacks. In addition to designing a secure system, predicted data on the model should be fed back into the pipeline to drive specific decisions [56]. If necessary, the deployed model can also be trained on newly available data to increase its performance.

Each phase of the AI lifecycle contains different complexities and potential for privacy breaches. For instance, during the model-building phase, developers can experiment with privacy-preserving mechanisms and observe their impact on model performance. However, once the model has been deployed, frequent changes may not be possible due to the risk of the model being exposed to attackers. Nonetheless, integrating necessary solutions into the development of AI systems across their entire life cycle is crucial to safeguarding privacy [26]. As such, privacy threats and solutions must be evaluated in the context of the AI life cycle.

IV. PRIVACY RISK CATEGORIZATION

Privacy standards and regulations are constantly shifting since AI as a technology continues to evolve and its applications continue to expand. We are also becoming increasingly aware of AI technologies' privacy implications and risks. Identifying and classifying the privacy risks related to the development of AI in light of existing privacy regulations can help mitigate these risks. A privacy risk source can be defined as an entity, process, or technology whose action may compromise data subjects' privacy, leading to intentional or unintentional privacy harm [57]. The action of the attacker or adversary may be due to unauthorized data processing through malicious intent [58]. The compromise may also be due to unauthorized data processing through malicious intent of an attacker or adversary, or due to unintentional incident, error or mistake [59]. Privacy risks can emerge in each phase of the AI life cycle and need to be proactively addressed. However, due to the rapid growth in AI-related applications, the design and deployment of AI algorithms are generally performed without adequate attention to governance, oversight, transparency, and accountability [60]. One of the main objectives of this work is to define a privacy risk framework to evaluate the vulnerabilities and risks related to AI development. Our proposed risk framework includes four categories, namely identification, making inaccurate decisions, lack of transparency in AI systems, and non-compliance with privacy regulations. Figure 3 presents the taxonomy of the identified privacy risks and their major contributing factors. These risk categories are further discussed in the following subsections.

In the context of protecting privacy in AI applications, it is important to consider the application domain as the context of

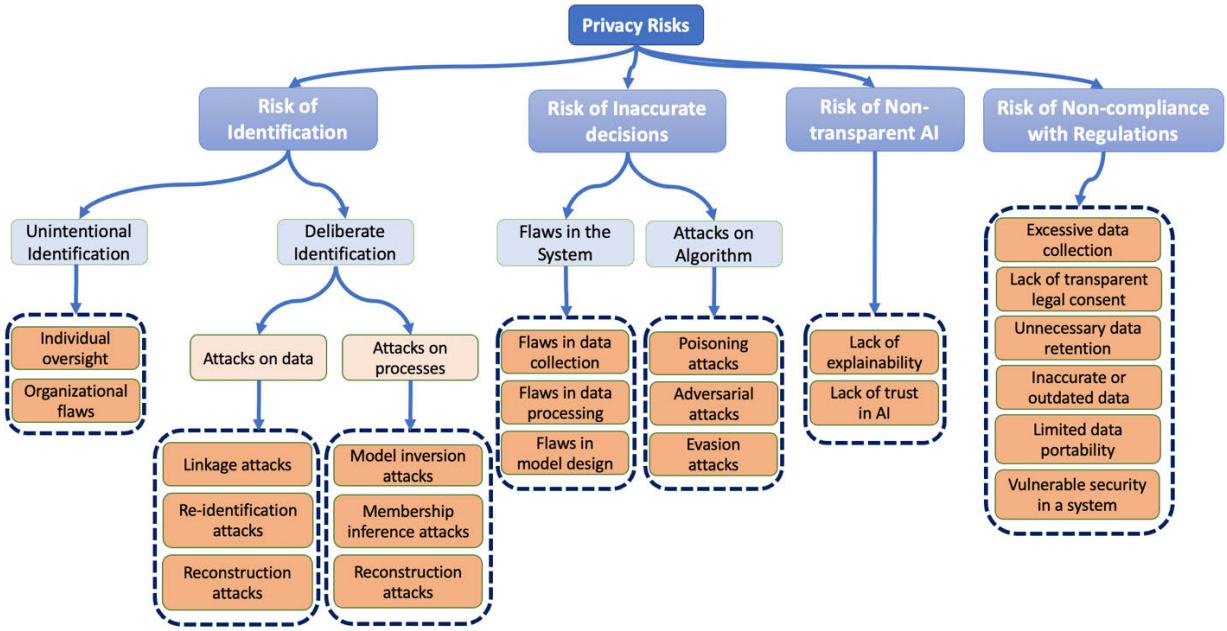

FIGURE 3. The Taxonomy of Privacy Risks.

application dictates the appropriate privacy approach. Different applications have different requirements and constraints, and the level of privacy protection needed may vary depending on the sensitivity of the data being used, the potential impact of a privacy breach, and the legal and ethical considerations surrounding the application. For instance, the privacy requirements for a healthcare AI application that involves sensitive patient data would be very different from those for a law enforcement application that involves public security concerns. In healthcare, privacy is a crucial ethical and legal requirement that must be protected, whereas in law enforcement, privacy considerations must be balanced with the need to protect public safety.

A. RISK OF IDENTIFICATION

The risk of identification refers to any action that may threaten to reveal the data subjects' identity. The identity of a data subject may be compromised by a malicious attack by an adversary [61] or by unintentional actions [62]. For example, insights generated from the analysis of datasets may lead to personal identity disclosure, which could occur without malicious intent. On the other hand, malicious attacks are more prevalent and require sophisticated approaches for mitigation [63]. Examples of malicious attacks include cybercriminals targeting databases to steal or infer personal information or PII about individuals to commit financial fraud.

1) UNINTENTIONAL IDENTIFICATION

Research by McAfee demonstrated that 43% of data breaches are caused by internal actors in an organization and that half of those breaches are unintentional [64]. In the context of AI systems, the analysis of personal data, whether anonymized or not, may lead to accidentally identifying a data subject. These unintended activities leading to identification can be further categorized into two:

- Individual oversight: An individual with access to personal or sensitive data may harm the identity of data subjects due to negligence. For instance, mailing sensitive data to incorrect recipients and transferring sensitive data to personal devices by employees are significant contributors to data breaches in healthcare [65]. This risk is exacerbated by the fact that mistakes made by individuals often go undetected until the disclosure of personal data has caused irreversible damage to data subjects.
- Organizational flaws: Unintentional leakage of personal data may also occur due to a lack of appropriate preventive measures by organizations. These measures can include technological solutions such as data leakage prevention systems [63] and company guidelines for dealing with sensitive data. The latter can also minimize employee oversight by promoting a culture of integrity in the organization by emphasizing the organization's responsibility towards privacy [66].

2) DELIBERATE IDENTIFICATION

This category deals with actions that lead to the deliberate identification of data subjects. The motives of adversaries vary from financial gains by external actors to corporate espionage by internal actors. In the context of AI systems, these attacks are risks that may take place deliberately on each

part of the system, i.e., input data, model, and output. Such attacks can target either data or processes to infer and misuse data subjects' identities.

a: ATTACKS ON DATA

Attacks on data can compromise data subjects' identity or their integrity by targeting databases, i.e., stationary data or data transfer channels. There are three prevalent forms of attacks on data that target personal identity:

- Linkage attacks: A database is de-anonymized using a linkage attack if the adversary using auxiliary information about a certain individual can reveal which record in the database corresponds to that individual [67]. Moreover, attackers often associate some available information with auxiliary data from various mediums, including the internet, public records, and domain knowledge [68], to identify individuals and their sensitive information.
- Re-identification attacks: In these attacks, anonymized or de-identified personal data can be matched with its owner [69]. The adversary may utilize different approaches to re-identify an individual, including linking datasets using background knowledge [70] and comparing mobility traces to infer sensitive information [71]. A study by De Montjoye et al. [36] has shown that it is possible to uniquely identify 95% of individuals by using human mobility data consisting of only four Spatio-temporal points. While individuals can be uniquely identified from such data, inferring their identity often requires the use of additional personally identifiable information that may not be available from the mobility data alone.
- Reconstruction attacks: These attacks involve partly reconstructing a private dataset by using publicly available information about the dataset. For instance, the adversary may reconstruct a probabilistic version of the original dataset used to train a model by using the model description as well as auxiliary information [72]. Moreover, Dinur and Nissim [73] demonstrated that adversaries may utilize some random queries to a database and combine the results of the queries to reconstruct sensitive data from a database.
- Derivation attacks: These attacks involve inferring sensitive information from non-sensitive data by exploiting correlations or patterns in the data. Consider the example of an e-commerce website that collects data on customers' purchases, such as the purchased items, product prices, and the frequency of purchases. An attacker could use this data to make educated guesses about a person's income level by analyzing the types of products they buy and their price range. For example, if a person frequently buys luxury items, high-end electronics, and expensive clothing, the attacker could infer that the person has a high income. In this case, the attacker is exploiting the correlation between the purchase history and the person's income level to derive sensitive information (the person's income) from non-sensitive data (the purchase history).

b: ATTACKS ON PROCESSES

Instead of exploiting data to identify an individual, adversaries may explicitly target AI algorithms. Generally, targets on the algorithm can be categorized into white-box attacks and black-box attacks. In white-box attacks, the adversary may possess some knowledge about the model or its original training data whereas in black-box attacks, the adversary does not have any information about the algorithm and is forced to probe the system to infer potential vulnerabilities [74]. Besides contributing to the risk of identification, attacks on processes can also harm individuals by changing the outcome of AI models (to be discussed in Section IV-B). There are three common attacks on processes that may result in identifiability:

- Model inversion attacks: In these attacks, an adversary gains access to an AI model to learn sensitive information about individuals [75]. An adversary can either infer the feature vectors or the general pattern of data used for model building through the deployed model. Fredrikson et al. [76] have shown that attackers can utilize a deployed ML model to recover recognizable images of people's faces given only their names. Model inversion usually occurs by the attacker submitting random input samples to a deployed classifier and observing the classification confidence for each input (black-box attack), intending to modify inputs to maximize the confidence values returned by the model.
- Membership inference attacks: In these attacks, an adversary utilizes various methods to reveal a user's membership to a dataset [77]. For instance, an adversary can generate a random record and run it by a deployed AI model to obtain predictions with confidence or probability values [25]. The adversary then continues to adjust the original record until a high probability value is obtained, in which case the curated record is almost identical to a member of the dataset. If attackers can ensure that an individual is a member of a given dataset, it is a positive membership attack. Likewise, when attackers can establish that an individual is not a member of a given dataset, it is referred to as a negative membership attack.
- Reconstruction attacks: In reconstruction attacks to processes, feature vectors used to design an AI model are available to the intruder, constituting a white-box attack. The attacker can use the extracted features to roll back or re-construct information about data subjects. Models that include explicit feature vectors of data samples, such as support vector machine (SVM) [78] and k-nearest neighbors (K-NN) [79], are more susceptible to reconstruction attacks. Combining extracted

features from the original data and the algorithm's parameters can lead to more severe reconstruction attacks.

B. RISK OF INACCURATE DECISIONS

Accuracy is among the core requirements for a trustworthy AI system. An inaccurate decision made by an AI algorithm can contribute to harmful social, political, and legal outcomes. For example, qualified applicants may be incorrectly rejected for employment or loans, or an innocent person may be arrested unfairly by an incorrect automated decision-making system. An inaccurate AI system can result from inadequate data collection, processing, and model design [80]. Inaccurate training data can impact AI performance significantly and have real-world consequences in various applications [81]. As an example, the presence of racial bias in the training set can potentially affect the relevance and accuracy of predictions for people of color/underrepresented groups [82]. Appropriate pre-processing is also important for improving the accuracy of several ML algorithms [83]. In terms of model design, an over fitted model can substantially impact model performance. A model is over fitting when it performs exceptionally well in training but fails to deliver generalized predictions on unseen data. Therefore, to ensure AI systems are accurate, it is necessary to address flaws in data collection, processing, and model design.

Additionally, attacks on databases or algorithms may compromise data integrity [84] and cause algorithmic biases, resulting in errors and inaccurate outcomes. Some prominent attacks that distort the results of the algorithms are as follows:

- Poisoning attacks: These attacks aim to distort the AI model's decision and impose biases on the outcome by contaminating the training dataset [85]. Therefore, these attacks are independent of the algorithm's structure, learning type, and lifelong learning capability, potentially impacting deployed models. Attackers usually perform these attacks on AI models that need re-training by injecting malicious samples during operation. Re-training is performed to keep the AI algorithms up to date. For instance, poisoning attacks can have devastating consequences in healthcare applications by causing the models to misdiagnose [86].
- Adversarial attacks: These attacks intend to deceive and manipulate AI models by injecting malicious inputs into the system [87], [88]. Usually, AI models are trained and tested on samples from the same statistical distribution. However, adversaries may inject data from different distributions or invalid sources to compromise the results by exploiting specific vulnerabilities [89]. Consequently, these attacks may impact the model's output and move the decision boundary resulting in inaccurate or biased decisions.
- Evasion attacks: In evasion attacks, an adversary aims to evade detection by obfuscating data content [90]. To this end, the adversary injects several instances with incorrect labels to train an AI model and alter its outcome [91]. Spoofing attacks against biometric verification systems [92] are an example of evasion attacks.

C. RISK OF NON-TRANSPARENT AI

The automated processing of personal data to analyze an individual's interests and personality is known as automated data profiling (ADP). The process of making a decision by automated means without human involvement, using factual or inferred data is known as automated decision making (ADM) [93]. Utilizing AI to attain decisions on credit scoring, hiring, and national security are examples of ADM that often include data profiling. In contrast, grading multiple choice questions using a pre-programmed application is an example of ADM that does not involve data profiling. According to many privacy legislation and recommendations, ADM and ADP should be transparent, interpretable, and free of bias and discrimination. According to OECD, transparency and explain ability are important principles for developing AI, and data subjects adversely affected by an AI should be allowed to challenge an AI-generated decision regarding them [94]. Interpretability means that decisions and predictions made by an automated system are understandable to humans [95].

Explainable AI (XAI) [96] is an active field of research promoting solutions toward increasing the interpretability of AI algorithms. In contrast to traditional ML algorithms, deep learning algorithms (that utilize deep neural networks) are more difficult to explain as they do not have a dedicated feature selection phase. The lack of explain ability of some deep learning algorithms is also known as the 'black box' phenomenon [97]. These algorithms sacrifice interpretability for accuracy, and by using complex non-linear associations and connections across the network, they become inherently uninterpretable to humans [98]. Moreover, it is unclear whether legislation such as the GDPR requires an explanation of ADM [99], [100]. Wagner [101] argues that due to frequent algorithmic changes and the algorithm being considered as intellectual property of an organization, it is not feasible to explicitly investigate the algorithms to determine interpretability. Rather, he maintains that important information about the development of algorithms, including the training data and what objectives the algorithms are optimizing for, can be revealed to the public. This aligns with the recently proposed Bill C-27 in Canada under the AI and Data Act, which requires developers to publish a plain-language description of the system, including how the system is intended to be used and the types of decisions it intends to make [102]. Lack of transparency in terms of how the AI system functions and makes decisions about individuals can lead to public distrust in AI. Therefore, prioritizing transparency by making the process of AI development open to the public can facilitate AI trustworthiness. While transparency and interpretability are important for promoting

trust and accountability in machine learning models, they can also potentially enable privacy attacks. Model explanations, used to enhance transparency and interpretability, can inadvertently reveal sensitive information about the training data or the model's decision-making process [103]. In particular, counterfactual explanations, designed to show how changing input features can impact the model's output, can be exploited to extract a target model's parameters and steal sensitive data [104]. Therefore, developers must carefully select strategies to enhance transparency in AI models without compromising privacy.

D. RISK OF NON-COMPLIANCE WITH PRIVACY REGULATIONS

Legislation and organizations such as GDPR [27], PIPEDA [30], CCPA [29], and OECD [28] have provided recommendations and policies on how to ensure the protection of individual privacy. Inadequate attempts to comply with privacy regulations and policy recommendations are a risk to privacy and can have significant consequences for organizations. Firstly, most privacy policies impose hefty sanctions on organizations for privacy-related infringements [27], [102], potentially causing financial harm in cases of violations. Moreover, not complying with regulations may hamper the business competitiveness of organizations as they would be considered unreliable to their competitors. Finally, non-compliance with privacy regulations can result in a loss of public trust in AI. Risks of non-compliance with privacy principles and best practices include:

- Excessive data collection: Excessive data collection in AI systems may compromise data subjects' privacy by increasing the risk of identification and exposing data subjects' behaviors. Furthermore, excessive data collection can result in additional harm in cases of data breaches or security attacks as the overall amount of data leakage is increased. Article 25 of GDPR states that appropriate technical measures must be taken to ensure that data collection and processing are limited to specific purposes [27]. Likewise, Principle 4 of PIPEDA's fair information obligates limiting the data collection only for the specific purpose and by fair and lawful means [30].
- Lack of transparent legal consent: Not informing the end-users about how their data is collected, stored, processed, shared, and disposed of can threaten their privacy. It is because ADM may utilize profiling to infer a data subject's behavior and presumed interests to provide recommendations. In this context, consent is necessary to ensure that the data subjects are aware of the decisions impacting them and have agreed to the use of their data in making these decisions [101]. Unambiguous, explicit, and comprehensive consent is also important for mitigating the risk of non-transparent AI. For example, websites collect cookies to facilitate users' access and interaction with the website. GDPR requires service providers to provide informed, specific, and freely given consent for collecting cookies [105] to ensure the data subject is aware of the service providers tracking practices and, in turn, ensuring the privacy of the data subject is not compromised. Consent is also required for collecting, processing, or disclosing personal information according to Principle 3 of PIPEDA [30]. Therefore, comprehensive and transparent legal consent is necessary for data collection in AI systems. However, such legal consent may not be applicable for AI systems in certain applications, such as surveillance and law enforcement.
- Unnecessary data retention: The accumulation of data in the AI system for an indefinite time may compromise privacy by increasing the risk of composition, linkage, and intersection attacks. Moreover, not retaining personal data long after its intended purpose will minimize the risk of the data being outdated and inaccurate, which will impact AI performance, and help the organization reduce security and storage costs. The right to erasure (Article 17) of the GDPR [27] obligates the data controller (e.g. service providers) to erase personal data when it is no longer required for the initial intended purposes.
- Incorrect or outdated data: Inaccurate and outdated data can potentially misrepresent an individual. Moreover,
- in the context of AI, it can lead them to generate inaccurate decisions about individuals, as discussed in Section IV-B. Therefore, data controllers must ensure personal information is kept up-to-date, complete, and accurate. The need for personal data accuracy is highlighted in Article 5 of GDPR [27] and Principle 6 of PIPEDA [30].
- Limited/lack of data portability: Data portability entails the ability of individuals to receive personal data concerning them and transmit those data to another data controller [27]. In many best practices and regulations, individuals have the right to receive their data in a structured, commonly used, and machine-readable format, i.e., interoperable format. The right to data portability, as defined by the GDPR, enables the development of effective privacy enhancement technologies [106]. The portability of personal data in AI systems allows data subjects to attain various AI service providers easily and promotes competition among these technology providers.
- Vulnerable security in systems: Ineffective security measures in an AI system can result in data breaches and disclosure of data subjects' information [107], constituting a critical privacy risk. Due to inadequate security measures, intruders can access the system to steal personal information. For example, through man-in-the-middle attacks, an outsider can access personal information without authorization [108]. Additionally, AI systems under reconstruction attack may compromise the model parameters, leading attackers to

reconstruct information about data subjects using the extracted features. The need for developing effecting security measures to protect sensitive information is highlighted in Principle 7 of PIPEDA [30]. Similarly, Article 32 of the GDPR obligates suitable technical and organizational methods to ensure the processing of data in a secure manner [27]. OECD also highlights security and safety as an essential component for trustworthy AI and that AI systems ''should be robust, secure and safe throughout their entire lifecycle'' [109]. The privacy challenges and solutions in light of the AI life cycle are discussed next.

V. PRIVACY RISK AND EXISTING SOLUTIONS IN THE AI LIFE CYCLE

A. PROJECT PLANNING

Planning is among the most integral phases of the AI life cycle that directly impacts all other phases and the outcome of the AI system. Not having well-defined objectives and requirements may lead to poor model performance, contributing to the risk of inaccuracy. Incomplete and incorrect requirements are significant contributors to the failure of a software project [110]. To this end, requirement elicitation is an integral step in the planning phase. Requirement elicitation is the process of identifying the necessary requirements of a system in collaboration with users, customers, and stakeholders [111] to implement project objectives successfully. The process of requirement elicitation ensures that developers have specific targets to meet and an appropriate timeline can be set for implementing each requirement, thereby enabling the successful implementation of project objectives. Moreover, it is necessary to include comprehensive documentation of the project plan, including the objectives, requirements, and limitations, to mitigate the risk of an inaccurate AI system. Developers and data scientists should collaborate with domain experts to plan data collection and model-building strategies. For instance, developing an AI system that can detect cancers from medical scans should involve oncologists and radiologists, the domain experts in this application. The project plan should also include strategies to de-identify and anonymize data, limit access to this highly-sensitive data, and privacy metrics to evaluate the protection of individual data. The project plan should identify evaluation methods to measure the effectiveness and accuracy of the AI system. Failure to select the correct evaluation metric and strategy will lead to performance and privacy-related issues after model deployment.

Additionally, a well-documented plan will improve the transparency of the AI system. It is also necessary to describe the system in non-technical and plain language based on the recommendations of various privacy policies [102]. To further increase trust and transparency in the AI system, other stakeholders, including the business team and sponsors, should be involved in the planning stage and agree on the vision and goals of the project [110]. This process will ensure all stakeholders agree on the necessary requirements and the privacy goals of the project. It is also essential to interview potential users and data subjects to understand and consider their privacy requirements at the planning stage. The requirement elicitation process should consciously embed features that can make the AI system more transparent. The transparency of an AI system can be enhanced by making the source code available among the stakeholders, and ensuring the project is well documented highlighting the data use, project requirements, and expected outcomes.

A limited understanding of the project requirements and poor planning may also increase the risk of identification. A retail company may, for example, analyze personal data, revealing insights about customers, which was not necessary for designing the AI model for forecasting company annual sales. Therefore, a thorough examination of the project's objectives and planned implementation strategy by utilizing the Privacy Impact Assessment (PIA) can mitigate these privacy risks to a great extent. PIA can establish that program managers and system owners have consciously integrated privacy protections throughout the development life cycle of a system [112]. PIA can help analyze the risks related to collecting, using, sharing, and maintaining sensitive information [113] and minimize privacy risks proactively. To be effective, PIA should be planned in light of the project's objectives, scope, and limitations [114]. In addition to PIA, the risk of identification can also be mitigated by preemptive security measures such as strong authentication, encryption, and access control strategies. Strong authentication and access control methods can prevent unauthorized access, and robust encryption methods can preserve confidentiality in case of illicit access. Besides introducing security measures, establishing privacy-focused ethics and guidelines on dealing with sensitive data can mitigate the risks of unintentional identification to a great extent. This includes organizing workshops to educate employees on best practices and common mistakes when dealing with sensitive data. Finally, the requirement elicitation process should also plan for effective anonymization and de-identification methods to ensure PII is removed from the collected data.

Integrating the recommendations and mandates of privacy regulations in the project requirements can alleviate the risks of non-compliance with privacy policies. These recommendations should be integrated proactively through privacy by design. The fundamental principles to be implemented throughout the project development include data minimization, transparent consent, limited data retention, data accuracy, secure data storing, interoperability, and transparency. The planning phase should also include strategies for developing consent management, data interoperability, effective system security, and methods for rectification or update of personal data. In addition, regulations and best practices recommend de-identification and anonymization of personal data at the source, i.e., irrevocably removing any PII and ensuring no trace back to identifiable information. To prevent AI systems from being attacked, identifying strategies

for extensive privacy and security testing is critical. Various privacy regulations also require the implementation of appropriate technical and organizational measures to guarantee privacy [27]. To this end, privacy metrics can help assess the protection and susceptibility of data in revealing private information [115]. Furthermore, privacy metrics can quantify the level of privacy acceptable by compliance officers and data subjects. The use of privacy metrics and PIA can also help demonstrate compliance with regulations. The planning phase should identify the necessary requirements for implementing PIA and privacy metrics. Finally, engagement of various user groups of the AI system, including data subjects and end users, through the AI development lifecycle will be instrumental to understanding and implementing their privacy needs.

B. DATA COLLECTION

This section outlines some privacy issues that may arise during data collection, transfer, and storage. Most of the privacy risks in this phase fall under the risk of non-compliance with privacy regulations. To alleviate the risk of excessive data collection, most privacy legislation recommend data minimization, i.e., personal data processing should be adequate, relevant, and limited to the intended purpose [116]. Personal data can be reduced after collection to further decrease the risk of excessive personal data processing. Attributes containing certain PII, including names and addresses, are not relevant to model training in most applications and therefore, should be removed or not collected in the first place. In applications where some PII, such as gender and age, are important learning features for the AI model, de-identification and anonymization methods should be applied. For instance, Goldsteen et al. [117] proposed reducing the granularity of input features by removing certain features or generalizing them. To handle the trade-off between reducing the collected data and imposing a minimum impact on the model's accuracy, they employed the knowledge encoded within the model to produce a generalization.

The integration of an appropriate consent management framework is necessary for data collection according to most privacy regulations. The collection and utilization of personal data without data subjects' consent is an invasion of their privacy. To facilitate this, Castelluccia et al. [118] recommend providing a data agreement, e.g., terms and conditions, for data subjects to have their consent prior to data collection. This agreement should elaborate on how personal data are collected, how they are transmitted and stored, why and how data are processed, and when data are deleted. The agreement should be clear, easy to understand, and accessible to data owners. It is also necessary to ensure that the data agreement is presented to the data subjects in a non-disruptive manner, without repeated requests for approval [118]. Data subjects should also have the right to accept or reject being subject to the data collection and processing. Furthermore, the consent should be part of a legal document known as a privacy policy [119] or privacy statement. This document should also include a privacy notice informing the data subjects when their data is collected and what data is collected. In addition to demonstrating compliance with privacy regulations, privacy policies also help with increasing AI transparency.

Retaining sensitive data for longer than necessary, i.e., data accumulation increases the risk of data loss through attacks. Moreover, the collected data may become outdated, impacting decisions about data subjects and contributing to the risk of inaccurate AI systems. Therefore, personal data may only be retained in the system until the contract terminates or data subjects are willing to keep their data for longer. Furthermore, companies should ensure that all available data are accurate and up to date, and individuals have the right to rectify their outdated or incorrect data [120]. Data controllers should introduce technical design and implementation [121] to facilitate the rectification or removal of specific data related to a data subject from all their servers and backups. A mobile or web interface can be introduced to enable data subjects to rectify or update their personal data. Interactive tutorials can be provided by the data subjects to ensure the interfaces are accessible and user-friendly for data subjects. It is also the responsibility of the data controllers to ensure third parties (other controllers and data processors) comply with data collection and data practices that have been established through consent, including data retention and rectification.

Under GDPR and other privacy regulations, data subjects have the right to receive a copy of their data, in a machine-readable format. To comply with the GDPR principle, an interoperable and standard data format [122] should be made accessible to data subjects. For example, Jaleel et al. [123] designed a framework to present medical data interoperability and standardization through the collaboration of healthcare devices. Interoperability also has other benefits in large systems that receive data from diverse sources, including data integration and processing. Moreover, data standardization, referred to as semantic interoperability, can help bring data into a commonly accepted format and definition that allows for data sharing, data integration, and collaborative research [124]. Bezuidenhout [125] investigated what infrastructures and resources for data standardization are needed to make data more accessible, interoperable, and reusable. Data controllers should provide suitable documentation and tutorial for data subjects to facilitate data portability. The documentation should explain how data subjects can access their data and have them transferred to another organization. Data controllers should also ensure safety measures when transferring data to data subjects or other platforms.

Safe data storage and transition are crucial to mitigate both the risk of an insecure system and the risk of identification. According to GDPR, data storage should comply with security and safety standards to protect data [126]. Also, GDPR mandates reporting personal data breaches to authorities in less than 72 hours. Attacks on data can occur during two phases; data-at-rest attacks occur on stored

databases, whereas data-in-motion attacks occur during the transition of data from one platform to another. Data linkage and re-identification attacks are examples that threaten data subjects' identification by targeting data-at-rest, while man-in-the-middle attacks target data-in-motion. According to Jain et al. [127] encryption techniques, including identity-based encryption, attribute-based encryption, and storage path encryption, are fundamental solutions for privacy protection during data storage. Additionally, a firewall between the storage server and the network [128] should be installed to enhance the privacy and security of the systems and defend against attacks. The firewall can control incoming and outgoing packets and filter out suspicious packets. Another preventive solution to manage data access is by using high-security protocols and access management. For example, a data catalog, which provides a structured listing of data assets in the available database to facilitate accessibility and security [129], can be used as a suitable protocol and access management tool. Data catalog uses metadata to help organizations manage their data and perform data governance by organizing data based on their importance. Furthermore, an appropriate identity and access management framework [130] is particularly important in collaborative projects, where multiple institutions access distributed resources. Such frameworks ensure personal data is only handled by authorized users. Local differential privacy [131] can be an effective solution in mitigating the risk of identification for various applications. In this context, individual perturbs their data before sharing it with the data collector, adding a small amount of random noise to their data in a way that preserves the overall statistical properties. For example, consider a survey asking people whether they have ever committed a crime. Each respondent would add a small amount of random noise to their answer before submitting it. This would make it difficult for the data collector to determine the exact response of any individual, but still provide insightful information about the overall population without compromising the privacy of individuals.

To safeguard against the risk of non-transparent AI, data controllers should integrate metadata to provide data integrity. The metadata should address how the data will be accessed, who has access to the data, what data is to be collected, and for how long will the data be stored. Providing metadata can also help data subjects demonstrate compliance with privacy regulations and authorities. In addition, meta-data can help interoperability by providing a standard format and definition for data transiting and processing. It is also the responsibility of the data controller to determine whether sensitive PII is part of the collected data and categorize any direct or indirect identifiers. Lack of transparency can also arise due to incomplete or non-existent consent of data subjects. Arnold et al. [132] proposed employing a comprehensive and transparent document or 'FactSheet' for AI systems to address transparency concerns. Such a document should contain sections on all relevant factors related to data privacy in AI systems, such as data collection, consent of data subjects, and intended use.

C. DATA PREPARATION

Ineffective data collection and technical faults, including faulty sensors and data loggers, may result in missing, distorted, or inaccurate values. Inaccurate and incomplete data can contribute to the risk of inaccuracy and impact decisions about data subjects. In cases when data is clean and free of incorrect values, the transformation of input features (e. g. normalization of numeric values) and labels (e. g. one-hot encoding) can improve model performance [133]. Privacy-preserving data cleaning can be utilized for interpolating missing data. For instance, Jagannathan and Wright [134] proposed a lazy decision-tree imputation method for data partitioned between two parties without revealing the computed model to either party, thus preserving privacy. Data refinement and statistical approaches to handling missing data [135] are other preprocessing techniques to mitigate the risk of inaccurate AI systems. Moreover, outliers and anomalous data can mislead the model and cause it to produce inaccurate outcomes. In this context, unsupervised ML approaches are popular for outlier detection. For example, the isolation forest algorithm can perform multivariate outlier detection by isolating anomalous data points and considering their relationship with other points in a tree structure [136]. Data visualization is another approach for outlier detection, enabling data scientists to analyze data distribution and determine the expected data range. Consequently, instances falling outside the expected range are denoted as outliers [137]. Furthermore, an illegitimate data point may be generated by an adversary [138] after the data collection phase. Such data points may be statistically similar to other legitimate points, making them difficult to isolate using the aforementioned approaches. In this case, generative adversarial-based networks are suitable for detecting such malicious data points by learning the adversarial features [139]. Detection and removal of incorrect values and outliers in the data preparation phase can improve the model's accuracy and generalization. Finally, feature engineering, which transforms raw data into meaningful representation using human expertise [140], is necessary to improve the performance of supervised ML algorithms. Examples of feature engineering include extracting information such as month and year from time-series data. In other cases, new features can be generated by mathematical transformations, including trigonometric transformation.

Developers can also integrate preemptive solutions in the data preparation phase to mitigate risks of personal data identification. In this context, the de-identification of personal data is necessary. De-identification refers to eliminating identifiable information, including names and phone numbers, from personal records to protect the privacy of data subjects. Similarly, pseudonymization is a preventive solution to preserve data privacy [141]. Pseudonymization replaces features that can identify a data subject with a value that does

not imply the data subject's identification, i.e., a pseudonym. Substituting student names with student numbers for course grade announcements is a simple example of pseudonymization. However, de-identification and pseudonymization alone are insufficient for providing privacy effectively and must be combined with other solutions, including dimensionality reduction. To this end, Jaidan et al. [142] demonstrated that using dimensionality reduction in privacy-preserving algorithms can decrease the risk of re-identification. Moreover, Principal Component Analysis (PCA) is an effective method for dimensionality reduction that provides linear transformation of features [143]. Besides PCA, other non-linear and autoencoder-based transformations [144] can be utilized to reduce dimensionality. These methods help anonymize data by transforming the raw data, which is more susceptible to identification, into a more complex representation. Linkage and re-identification are attacks on data, whereas reconstruction attacks are privacy threats on extracted features in the preprocessing phase. Reconstruction attacks may target the available feature vectors in the dataset or during model building. Naehrig et al. [145] proposed cryptography models such as homomorphic encryption to mitigate the risk of reconstruction attacks. Homomorphic encryption uses a public key for data encryption and an algebraic system to work with encrypted data. The main advantage of homomorphic encryption is that AI models can be developed with encrypted data without decryption. This method can ensure data privacy because only data controllers with the matching private key can decrypt data when needed. To mitigate the risk of adversarial attacks, adversarial feature desensitization using generative adversarial networks (GANs) [146] should be applied during feature engineering. Natural and adversarial data cannot be discriminated if the learned features are invariant towards adversarial perturbations.

D. MODEL DEVELOPMENT

Developing an accurate AI model is one of the main objectives during the model design phase. Developers can make changes to an AI system iteratively and evaluate performance until the desired outcome is obtained. For instance, an interactive model and human-in-the-loop (HITL) system [147] can help developers track changes and determine the best-performing conditions. HITL leverages the power of AI and human intelligence to optimize AI models. To this end, an expert supervises training, tuning, and testing tasks, especially in edge points where the algorithm has low confidence in decisions or encounters a problem. Xin et al. [148] proposed a Helix HITL system that is fast and effective for improving model accuracy. Responsive feedback and automation are the main advantages of such systems. In addition, using appropriate evaluation metrics to measure model performance is essential to mitigate the risk of inaccurate AI systems. For instance, in imbalance classification problems, commonly used metrics such as classification accuracy may be influenced by the majority class [149] and thus fail to represent model performance accurately. In this context, AI developers must be aware of various evaluation metrics and their limitations as it applies to a given problem. It is also necessary to use evaluation strategies that minimize prediction errors. For instance, training error and standard k-fold cross-validation contain biases that must be adjusted [150]. Therefore, appropriate evaluation metrics and strategies should be selected in collaboration with data scientists and AI researchers. Moreover, determining the optimal model parameters, i.e., hyperparameter tuning, can improve model performance [151]. In this context, developers should experiment with various strategies, including grid search and Bayesian optimization [152], to determine the set of optimal model parameters. Overfitting [153] is a significant challenge in training deep learning algorithms and can hamper model performance. A model overfits when it performs exceptionally well on the training set but fails to generalize to unseen data. Developers should integrate various techniques, including dropouts of weights [154] and regularization [155], to avoid overfitting.

The lack of interpretability in some AI models contributes to the risk of non-transparency [156]. Utilizing inherently interpretable algorithms, such as decision trees, logistic regression, and linear regression, that include meaningful parameters to explain their predictions [157] can improve model transparency. Certain properties of a model, including linearity, monotonicity, and interaction, can help explain some of the results. In monotonic models, such as logistic regression and one-layer neural networks, the relationship between input features and target outcomes is correlated; therefore, mapping input features to the target can explain the decision-making procedure. Although it may be feasible to track the interactions between the input and output of the model in simple problems, the traceability of the model decreases for more complex models. On the other hand, in black-box models, i.e., algorithms with many parameters such as deep neural networks, it is not feasible to track processes and interpret the reason for a decision. Several post-hoc interpretation algorithms have been developed [158] to achieve a level of interpretability in black-box models. Post-hoc interpretation algorithms and XAI can convert black-box algorithms into 'glass-box' by adding interpretability to the AI models [98]. For example, Ribeiro et al. [159] proposed an interpretable and model-agnostic explanation of classifiers by learning model explainability locally around the predictions and framing submodular optimizations. Also, Lundberg and Lee [160] provided model interpretability by assigning feature importance values for each prediction. Integrating such techniques during model development can increase AI transparency. Predictions made by tree-based models can also be explained by the Shapley additive explanation (SHAP) framework [161]. SHAP computes the contribution of each feature in making a prediction. The Shapley values are computed based on game theory, and they indicate the impact of the features on a prediction.

The presence of bias and discrimination in an AI model [162] contributes to the risk of inaccurate AI. Bias and discrimination also hinder the trustworthiness of AI models and contribute to the risk of non-transparency. Although bias and discrimination can be minimized by appropriate data collection, an ineffective model development may also contribute to AI bias. A three-stage approach to managing bias in AI systems was proposed by [42] that includes managing bias during pre-design, model development, and deployment of an AI system. The study also recommended engaging the stakeholders to provide feedback to mitigate the risk of bias in AI systems. Moreover, in imbalanced classification problems, there is a potential for bias toward the majority class. To overcome this problem, oversampling minority classes and cost-sensitive learning are effective solutions [163]. Other solutions to the class imbalance problem include data augmentation [164] and algorithmic level modifications [165]. Bias in classifiers can also result from multi-modal datasets. In this context, Gat et el. [166] proposed a regularization approach based on functional entropy to promote equal contributions from each modality.

The risk of identification needs to be addressed proactively during the model development phase. The most prevalent reconstruction attacks utilize feature vectors in AI models. Therefore, algorithms that store explicit feature vectors, including SVM and K-NN, are more susceptible to these attacks [167]. Reconstruction attacks can be mitigated by encoding the feature vectors before storage. For instance, Haghighat et al. [168] demonstrated that the encryption of facial features in a biometric database minimizes information leakage without compromising model performance. Some attacks, such as membership inference and model inversion, rely on model prediction output and the outcome of the algorithms, i.e., class labels in classification or predicted values in regression methods. The effectiveness of such attacks can be minimized by limiting the intruder's knowledge about the system's results [89]. For instance, reporting only the predicted class labels of a classifier instead of probability values make it difficult for adversaries to perform these attacks. Moreover, differential privacy can prevent several attacks on AI models, including linkage and reconstruction attacks [169]. Various forms of differential privacy have been developed to reinforce AI models against adversarial attacks. For example, Agrawal and Srikant [170] demonstrated that adding random noise to the input data through differential privacy can resist attacks on AI systems. Similarly, Kim and Winkler [171] randomized data by multiplying noise with a known statistical distribution. The reconstruction of the original values is more difficult in the noise multiplication approach, making it suitable for preserving privacy. Despite many benefits of differential privacy, Ding et al. [172] demonstrated that there is no guarantee to confront the attacks by these methods comprehensively. Therefore, it is necessary to run a candidate algorithm numerous times and test it to detect violations of a specific differential privacy algorithm.

As previously discussed, data under homomorphic encryption can be used in building a model without decryption, providing complementary assurance in privacy-preserving methods. In collaborative applications, when different stakeholders want to attain a common objective but not share their data, training an AI algorithm is difficult due to limited data availability. For example, hospitals may only have a limited number of patient data for a specific medical application and thus require collaboration with other hospitals to obtain meaningful results. In this context, decentralized processing methods like federated learning [173] are instrumental in addressing privacy concerns while allowing collaborative learning. Decentralized learning approaches can provide secure multi-party computation, shared data storage, and a high level of privacy [174]. In the case of a central learning model, datasets can remain locally in each device and do not need to be transferred between edges. Variations of federated learning, including horizontal, vertical, and transfer learning, have been proposed to address privacy and security concerns [175]. Despite many benefits of federated learning, data locality during the training process cannot guarantee the privacy of centralized algorithms. Consequently, for comprehensive protection against adversarial attacks, it is necessary to integrate different types of protective algorithms.

Since model evaluation is an important subphase of model development, some privacy solutions should be specifically considered during this stage. The developers should ensure that the model is tested with representative data to avoid any potential risk of inaccuracy. When a dataset is partitioned into validation and testing sets, a random split may cause data from specific classes (in classification) or distribution to be excluded in a subset. To avoid such sampling biases, strategies such as stratified sampling [176] and stratified cross-validation [177] should be used. Moreover, XAI-based approaches remain important solutions in this subphase to mitigate risks of non-transparency and non-compliance. However, some XAI methods can be privacy-invasive, and therefore, XAI methods should be selected with privacy in mind. Haque et al. [178] highlight that trust, transparency, understandability, usability, and fairness are the significant XAI factors that impact AI adoption and use.

E. MODEL DEPLOYMENT

The design and development of a secure and privacy-enhanced AI system do not completely guarantee its safety post-deployment. Therefore, preemptive measures should be taken to mitigate the risk of identification after deployment. The outputs of a deployed model may reveal excessive information about the original dataset [179]. Consequently, an adversary can query a deployed model and infer sensitive information related to the data (model inversion and membership inference attacks) used for model building. The deployed model should reveal as little information as possible to prevent such threats. For instance, a classification model should only provide predictions of classes and not probability

values. However, embedding privacy requirements in a system lead to other restrictions on the system. For example, adding noise to data or features may impact the model's performance, whereas adding interpretability may increase model complexity. As such, there is a trade-off between the level of privacy and the model's performance and complexity. For instance, a facial recognition model trained on a large dataset of photos may be more accurate than a model trained on a smaller dataset, but the larger dataset could also contain sensitive information such as people's identities and locations. Additionally, the security measures implemented pre-deployment should be audited periodically to minimize attacks. The developers should perform regular security updates and integrate the latest security features into the system.

The deployed system should be safeguarded against intrusion that changes the system's outcome, leading to the risk of inaccuracy. Several attacks, including poisoning, adversarial, and evasion, inject distorted data into the deployed models to alter model outcomes. Handling extensive attacks on AI systems is impossible without taking intelligent defensive actions. Therefore, frameworks, such as cyber threat intelligence [180], should be integrated into the AI system. In addition to using intelligent defensive systems, educating the system's end-users and making them aware of probable risks can reduce unintentional privacy issues and prevent several attacks on the system. Furthermore, it is also necessary to retrain the model with time to keep the outcomes relevant.

The statistical properties of data may change with time, and consequently, the model performance may decrease. In this context, an automatic retraining algorithm [181] can facilitate keeping the model up to date. The retraining process should consider necessary security measures so that adversaries cannot inject malicious data. For example, retraining the model online may potentially expose the parameters and training data to adversaries. On the contrary, it is wiser to perform the retraining offline and re-deploy the model after training.

The developers should continue to engage with data subjects post-deployment to understand their privacy requirements. The developers should also consider data subjects' privacy expectations and include data subjects in testing the deployed model. The AI system should be extensively evaluated for performance, transparency, and security using effective software testing methods. This includes testing the deployed AI system using Alpha testing, Beta testing, and user acceptance test [182] by engaging the end-users and stakeholders.

Moreover, privacy policies and other legal documents should be accessible to data subjects and should change with the data and technology practices of the organization. Some privacy commissions provide privacy policy templates [183] to organizations to enable them to write complete and readable privacy policies. An appropriate notice mechanism should be implemented to inform data subjects about system changes. To this end, Audich et al. [184] recommended the automatic categorization of privacy policies using NLP. Due to the complexity of privacy policies in terms of their length and language, NLP tools allow data subjects to engage with privacy policies conveniently. The deployed model should also be periodically audited for privacy issues to increase trustworthiness and demonstrate compliance with regulations. It is necessary to ensure that the auditors are external to avoid potential conflicts of interest. The auditing system should be comprehensive, allowing auditors to monitor data, data distribution, and AI system performance.

The privacy risks and their solutions are categorized in Table 1 by the different phases of the AI life cycle.

VI. DISCUSSION AND FUTURE RESEARCH DIRECTIONS

Although privacy legislation such as GDPR and organizations such as OECD aims to recommend best practices to address privacy concerns in AI systems, many complex applications and requirements of AI have not been fully considered. For example, autonomous vehicles integrate sensory technologies and AI algorithms to make more accurate real-time decisions [185]. However, decisions made by AI algorithms in such cases require analyzing the benefits and risks that are not straightforward. For instance, the AI algorithm may decide to cause a minor accident to avoid a potentially significant

collision with other vehicles. In such cases, it is challenging to recognize AI behavior, ethics, and policies and define accountability in regulations. Thus, establishing rules to harness AI applications in an ethical and privacy-preserving manner is more complicated and requires further attention and a nuanced approach. Considering the emerging underlying changes in AI technology, evaluating and updating the rules regularly alongside the changes in processes is needed. However, due to the evolving nature of AI, newer algorithms may not be fully comprehended by lawmakers. Regulatory bodies should therefore involve AI researchers and scientists along with legal experts to present emerging AI algorithms in a more non-technical approach.

In addition, standards and risk management frameworks, such as the ISO/IEC 23894:2023,[1] provide direction on managing risks associated with the development, deployment, and utilization of AI. These standards can provide risk assessment and risk treatment, including the identification of potential threats, vulnerabilities, and impacts on privacy and security. Moreover, they can enable developers to communicate privacy risks to stakeholders, including users, customers, and regulators.

The emergence of generative AI, such as large language models, poses significant privacy and trust risks. These models are capable of generating highly realistic text, images, and videos, which can create convincing deepfakes, impersonate individuals, and spread misinformation. This can have serious consequences for individuals and organizations, including reputational damage and financial losses. Furthermore, the

[1] https://www.iso.org/standard/77304.html

TABLE 1. Privacy Risks and Solutions in the AI Life Cycle.

AI life cycle phase	Privacy risks	Solutions
Project Planning	Risk of inaccurate AI systems	Comprehensive documentation of project plan. Requirement elicitation [111]. Identify appropriate evaluation metrics and strategies. Collaborate with domain experts.
	Risk of non-transparent AI	Document project in plain language [102]. Engage stakeholders and interview data subjects. Ensure data use, project requirements, and expected outcomes are highlighted.
	Risk of identification	Privacy impact assessment [112], [114]. Preemptive security measures such as strong authentication, access control, and encryption. Establish privacy-focused ethics and guidelines on dealing with sensitive data. Plan for effective anonymization and de-identification methods to ensure PII is removed from the collected data.
	Risk of non-compliance	Integrate recommended principles like data minimization, transparent consent, limited data retention, interoperability, and transparency throughout the AI development. Identify strategies for extensive privacy and security testing. Identify project impact assessment and privacy metrics [115]. Engage various user groups, including data subjects and end users, throughout the AI development.
Data Collection	Risk of non-transparent AI	Include metadata with relevant data collection information. Provide comprehensive and transparent document or 'FactSheet' describing data collection and processing [132], [118], [119].
	Risk of identification	Identify PII and non-PII elements. Embed firewall in the system [128]. Manage data access by high-secure protocols, encryption, and access management [127], [130]. Local differential privacy [131].
	Risk of non-compliance	Remove PII from the dataset not relevant to model training. De-identification and anonymization at source for other PII. Provide comprehensive and transparent data consent [118]. Introduce a mobile or web interface for subjects to maintain data accuracy. Provide data interoperability and standardization [122], [123]. Embed firewall, encryption, and security in system [127].
Data Preparation	Risk of inaccurate AI systems	Data transformation including normalization and encoding [133]. Privacy-preserving data imputation and outlier detection [134], [136]. Feature engineering using domain knowledge and transformation [140].
	Risk of identification	De-identification of PII and pseudonymization [141]. Dimensionality reduction using PCA [143] and autoencoders [144]. Homomorphic encryption [145]. Adversarial feature desensitization [146] using GANs.

TABLE 1. *(Continued.)* **Privacy Risks and Solutions in the AI Life Cycle.**

Model Development	Risk of inaccurate AI systems	Introduce iterative approaches like human-in-the-loop [147]. Utilize appropriate evaluation metrics and strategies [150]. Remove model bias by engaging stakeholders. Resampling of data and cost-sensitive learning to remove bias in class imbalance problems [163].
	Risk of non-transparent AI	Use interpretable algorithms such as decision trees and linear regression [157]. Include Post-hoc interpretation algorithms [158] and XAI concepts in model design [98]. SHAP framework for tree-based models [161]. Remove model bias by engaging stakeholders.
	Risk of identification	Encrypt features to minimize information leakage [168]. Limit intruder's knowledge to system by providing minimal output. Integrate a version of differential privacy [169], [170], [171]. Utilize decentralized learning methods including federated learning [173].
Model Deployment	Risk of inaccurate AI systems	Introduce intelligent defensive systems to avoid poisoning and other attacks that impact performance [180]. Introduce model retraining with time and ensure offline retraining [181].
	Risk of non-transparent AI	Continue to engage system users and stakeholders. Make privacy statements conveniently available and reduce complexity using NLP [184]. Implement a notice mechanism to inform data subjects about system changes.
	Risk of identification	Introduce intelligent defensive systems such as cyber threat intelligence [180]. Educate the system's end-users to avoid compromising privacy. Test the deployed system using Alpha, Beta, and user acceptance testing [182].
	Risk of non-compliance	Introduce comprehensive audit system, allowing auditors to monitor data, data distribution, and AI system performance. Ensure the audit is performed by external auditors.

data used to train these models can also contain sensitive information, such as personal or financial data, which can be exposed if the models are not adequately secured. It is thus necessary to implement robust data protection measures, such as data anonymization and encryption, and to ensure that access to the models is restricted to authorized individuals. Additionally, transparency and explainability are crucial for establishing trust in these models, as users need to understand how the models work. Privacy risks and mitigation of generative AI needs specific research attention to help identify and address the privacy and trust risks associated with generative AI.

While all the stages in AI development remain significant, project planning is perhaps the most important; *failing to plan is planning to fail!* The planning stage sets the blueprint for data handling throughout the project. Even though the direct risk of privacy breach is low at this stage since no data has been collected yet, the decisions made during this phase have a substantial impact on privacy risks in later stages. Neglecting relevant privacy regulations or failing to anticipate the need for PETs can set the stage for significant privacy risks in the latter phases. During data collection, without proper plans set earlier, privacy risks can emerge in the form of collecting more data than necessary, failing to adequately de-identify data, or not obtaining consent from the data subjects. The data preparation stage can compound these risks, as poor handling of sensitive information, poor anonymization, or biases can lead to inaccurate decisions. In the model development phase, the use of certain algorithms could cause explainability and trust issues, potentially leading to biased or incorrect models. Therefore, strict adherence to best practices, industry guidelines, and meticulous documentation of project planning can notably reduce privacy risks in AI development. Amidst the global proliferation of AI systems, researchers must prioritize all stages of AI development, particularly data collection and preparation due to their substantial privacy risks that require technical solutions. By refining data collection and preparation techniques,

researchers can minimize privacy breaches while ensuring effective AI development. Introducing clear standards and best practices for these stages can guide responsible data handling. Furthermore, it is vital to study other privacy concerns related to AI model interpretability, the potential for AI systems to be used in ways that infringe on privacy, and the legal and ethical considerations of AI.

A. IMPACT OF PRIVACY REGULATIONS ON AI DEVELOPMENT

The introduction of various privacy legislation and best practices or recommendations has impacted the development of technology and the deployment of websites and applications [186]. This section summarizes the potential impacts of privacy legislation on AI development.

- Most policies recommend a privacy-by-design or privacy-first approach to technological development. Thus, developers are required to possess a comprehensive knowledge of the AI lifecycle and apply a software engineering approach to developing AI applications.
- The timeframe required for AI developers to obtain the necessary data to develop AI applications may be significantly longer, mainly due to the various requirements and restrictions by privacy legislation in data collection. Some of the requirements include comprehensive and transparent consent and data minimization. Moreover, integrating privacy-preserving techniques throughout the AI lifecycle and extensive testing to guarantee privacy increases the development time. Therefore, the completion of any AI project will require a longer timeframe.
- Integrating privacy-enhancing technologies into AI development may also require technical expertise in other domains or experts with interdisciplinary skills. For instance, safeguarding against the risk of identification requires collaboration with cybersecurity experts in implementing technologies such as access control and firewall. Consequently, this increases the developmental cost of AI, making AI applications more expensive.
- Privacy regulations and best practices are continuously evolving, and some policies apply nationally or regionally while others on a global scale. Therefore, AI developers should collaborate with legal experts to fully comprehend and abide by privacy regulations. Similarly, developers should collaborate with external auditors to demonstrate compliance with privacy regulations. These collaborations increase the timeframe required for project completion along with the costs.
- PIA and regular privacy audits post-deployment are necessary to guarantee the privacy of AI applications. However, these requirements restrict developers from deploying AI applications. For instance, developers should design and deploy AI algorithms to facilitate audits post-deployment without interrupting the application service. Therefore, developer's need to have advanced software engineering skills to fulfill the requirements of AI deployment.

B. FUTURE RESEARCH DIRECTIONS

The privacy solutions discussed in the previous section contain several limitations that should be addressed in future research. There is also a need to develop new solutions that apply to specific phases of the AI lifecycle. This section highlights open research challenges to mitigate privacy risks in the AI lifecycle.

Privacy metrics are an essential tool in the planning phase of the AI lifecycle for improving the safety and reliability of the AI system by evaluating the susceptibility of AI models in disclosing PII. Moreover, privacy metrics can help demonstrate compliance with privacy legislation. However, the lack of standard privacy metrics for AI systems is an open challenge that should be addressed in future research. The development of standard privacy metrics will enable researchers to obtain benchmarks on the adequate level of privacy required in AI development. The development of privacy metrics should be in collaboration with research scientists, policymakers, and AI developers. Privacy metrics should be comprehensive and cover various AI applications, including supervised learning, unsupervised learning, and reinforcement learning. Privacy metrics are also relevant in the development and deployment phases of the AI lifecycle.

Data minimization is essential for preserving privacy in the data collection phase. However, it is challenging to determine the minimal amount of data required to build the AI model while reaching the desired outcomes. Therefore, future research needs to investigate the optimal amount of data for training an AI model without compromising privacy. Introducing a framework to determine the required data for developing effective AI algorithms would allow developers to demonstrate compliance with regulations and enable officers to hold accountable organizations that fail to minimize data. Moreover, comprehensive and transparent consent is necessary for personal data collection. In many cases, consent and privacy notice are presented in a lengthy and complex format for data subjects to comprehend. Therefore, a standard and concise consent form should be introduced by researchers for data collection. The performance of an AI system may be compromised when the available data is insufficient. In such cases, generating synthetic data can help the AI model learn the necessary parameters. Although there are various algorithms for generating synthetic data, there is a need to introduce a privacy-preserving model for synthetic data generation.

The performance and privacy levels of the AI system can be significantly improved in the data preparation phase. Specifically, in larger systems, a unified and interoperable framework allows receiving data from diverse sources to improve model performance. However, the lack of a reliable protocol for providing interoperable systems remains a research challenge. Consequently, future research should

focus on data portability and interoperability of AI systems to improve collaborative research. Moreover, various data-cleaning approaches are available for numeric and categorical data to improve model performance. However, privacy-preserving data cleaning approaches for time series, audio, and image data need to be investigated in future research. This includes developing a suitable transformation of these data formats to enhance performance and maximize privacy.

In the model development phase, detecting and managing bias and discrimination in AI applications is crucial to mitigate the risks of inaccuracy and non-transparency. However, detecting and eliminating bias in AI systems have not been sufficiently addressed in the existing literature. Therefore, researchers need to collaborate in setting up guidelines and recommendations to mitigate bias in various forms of AI. Moreover, handling the trade-off between the final model's accuracy, complexity, and interpretability should be addressed in future research to increase AI trust. Furthermore, decentralized learning approaches are essential in providing collaborative learning securely. Differential privacy can also be integrated into federated learning systems for enhanced privacy [187]. However, communication bottlenecks of federated learning and the inference attack over exchanged messages during the training phase are still limitations of these methods that need to be addressed in future research. Additionally, blockchain-based solutions enhance privacy, security, and performance [188], [189], [190] and should be investigated for privacy guarantees in decentralized data processing. For instance, Freund et al. [191] discussed the influence of different phases of the data lifecycle on the compliance of the GDPR principles in the treatment of data using blockchain technology. Such analysis should also be conducted from an AI lifecycle perspective.

Privacy is an essential consideration for AI applications running on resource-constrained devices like microcontrollers and wearables. These devices often collect sensitive data about their users, such as biometric data or location information, and transmit it over networks with limited security capabilities. Without proper privacy protections, this data could be vulnerable to interception and misuse with potentially harmful outcomes. More complex privacy solutions may not be suitable for model training on these devices due to their limited computation power. Therefore, investigating privacy-preserving techniques in AI applications for resource-constrained devices is an important research area to safeguard the privacy and security of users' data.

There is also a need for the scientific community, developers, government, and the general public to collaborate in advancing our understanding of trustworthy AI system, standardization of ethical and trustworthy concepts, and metrics to measure and evaluate those concepts. A framework should be developed to enhance collaboration in the AI research community for knowledge transfer and sharing research contributions on emerging and novel AI system attacks and mitigation strategies. In addition to increasing AI transparency, such collaborations would also positively impact research in AI development.

Auditing and PIA are essential strategies for assessing the level of privacy protection. However, existing compliance tools and PIA provides a general measure of privacy in a system. Since AI systems are more complex and contain different phases of development with a greater possibility of privacy breaches, there is a need for novel auditing and compliance tools for privacy-preserving AI systems. Moreover, interoperability in AI systems is necessary to provide data portability to data subjects. Consequently, a platform approach [192] is needed to enhance semantic, operational, and legal interoperability in AI systems.

It is necessary to evaluate the level of privacy in AI systems periodically after deployment to address existing vulnerabilities and emerging threats. External auditors can continue to assess the privacy risks in deployed models. However, the lack of a standard protocol for deploying AI systems makes it challenging for auditors to access deployed models for privacy evaluation. In many cases, the auditors may need technical expertise to perform extensive tests on deployed models. Therefore, to facilitate audits and post-deployment privacy impact assessments, there is a need for a standard protocol for deploying AI systems. For instance, it is necessary to ensure that only the auditors have access to the deployed model for assessment and that deployed AI systems are audited offline to avoid malicious attacks.

Following are the most notable future research directions:

- Develop comprehensive privacy metrics for assessing vulnerabilities in AI systems.
- Introduce a standard and concise consent form for data collection.
- Explore a privacy-preserving mechanism for synthetic data generation.
- Focus on data portability and interoperability of AI systems to improve collaborative research.
- Develop guidelines and recommendations to mitigate bias and inaccuracy in AI.
- Explore blockchain-based machine learning solutions for increased privacy, security, and performance.
- Investigate privacy-preserving techniques for resource-constrained devices.
- Develop a standard protocol for deploying AI systems to facilitate audits and post-deployment privacy impact assessments.

C. RESEARCH IMPACTS

This paper highlighted the need for privacy in developing AI algorithms and discussed state-of-the-art solutions to mitigate privacy risks that apply to various stages of AI development. There are several important implications of this paper. First, it sheds light on AI as a lifecycle or process and highlights the need to approach privacy in the context of the AI lifecycle. This approach complements the privacy-by-design concept

and allows AI developers to design algorithms with privacy embedded as a core component. The breakdown of AI development into key phases enables developers to integrate privacy-preserving solutions more comprehensively. In many cases, there are technical barriers between AI researchers and lawmakers due to the emerging and complex nature of AI algorithms. Therefore, introducing a framework to categorize relevant privacy risks facilitates collaboration between privacy lawmakers and AI developers. The paper also discusses state-of-the-art solutions to various privacy risks associated with each phase of the AI lifecycle, which enables AI developers to identify and leverage existing privacy solutions in their development. Furthermore, the potential impacts of privacy legislation on AI development were discussed in this paper, enabling AI projects to manage their resources efficiently. Finally, identifying existing research gaps allows researchers and scientists to focus on developing new technologies to mitigate various privacy risks in the AI life cycle.

VII. CONCLUSION

AI algorithms can help organizations, industries, and governments to improve core business processes and make better decisions. However, due to the data-driven nature of AI, privacy preservation in AI is more complex than traditional data privacy protection. Compromising privacy in each stage of an AI project can influence the entire system. As a result, this survey investigated privacy challenges throughout the AI life cycle. To this end, the privacy risks were examined in five AI life cycle phases: planning, data collection, data preparation, model building, and deployment. The paper also introduced a framework to classify privacy risks into four categories: risks of identification, inaccuracy, non-transparency, and lack of compliance. The privacy-enhancing technology and solutions were discussed to address the risk categories in the context of each phase of the AI life cycle. The paper also discussed the implications of the survey and the impacts of privacy regulations on AI development. Open challenges and research gaps were also identified, including the need for standard privacy metrics and interoperability in AI systems.

REFERENCES

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. London, U.K.: Pearson, 2010.

[2] A. M. Turing, "I.—Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, Oct. 1950, doi: 10.1093/mind/LIX.236.433.

[3] Q. Lang, C. Zhong, Z. Liang, Y. Zhang, B. Wu, F. Xu, L. Cong, S. Wu, and Y. Tian, "Six application scenarios of artificial intelligence in the precise diagnosis and treatment of liver cancer," *Artif. Intell. Rev.*, vol. 54, no. 7, pp. 5307–5346, Oct. 2021, doi: 10.1007/s10462-021-10023-1.

[4] K. Olorunnimbe and H. Viktor, "Deep learning in the stock market—A systematic survey of practice, backtesting, and applications," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 2057–2109, Jun. 2022 doi: 10.1007/s10462-022-10226-0.

[5] Y. Peng, E. Liu, S. Peng, Q. Chen, D. Li, and D. Lian, "Using artificial intelligence technology to fight COVID-19: A review," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4941–4977, Aug. 2022, doi: 10.1007/s10462-021-10106-z.

[6] M. Ruckenstein and J. Granroth, "Algorithms, advertising and the intimacy of surveillance," *J. Cult. Econ.*, vol. 13, no. 1, pp. 12–24, Jan. 2020.

[7] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111–125.

[8] A. Cavoukian, A. Fisher, S. Killen, and D. A. Hoffman, "Remote home health care technologies: How to ensure privacy? Build it in: Privacy by design," *Identity Inf. Soc.*, vol. 3, no. 2, pp. 363–378, Aug. 2010, doi: 10.1007/s12394-010-0054-y.

[9] S. Spiekermann and L. F. Cranor, "Engineering privacy," *IEEE Trans. Softw. Eng.*, vol. 35, no. 1, pp. 67–82, Jan. 2009, doi: 10.1109/TSE.2008.88.

[10] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individual Differences*, vol. 103, Apr. 2023, Art. no. 102274, doi: 10.1016/j.lindif.2023.102274.

[11] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–36, Mar. 2022.

[12] A. Oseni, N. Moustafa, H. Janicke, P. Liu, Z. Tari, and A. Vasilakos, "Security and privacy for artificial intelligence: Opportunities and challenges," 2021, *arXiv:2102.04661*.

[13] A. Boulemtafes, A. Derhab, and Y. Challal, "A review of privacy-preserving techniques for deep learning," *Neurocomputing*, vol. 384, pp. 21–45, Apr. 2020, doi: 10.1016/j.neucom.2019.11.041.

[14] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos, "Privacy and security issues in deep learning: A survey," *IEEE Access*, vol. 9, pp. 4566–4593, 2021, doi: 10.1109/ACCESS.2020.3045078.

[15] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 1–39, Jun. 2022.

[16] C. S. Wickramasinghe, D. L. Marino, J. Grandio, and M. Manic, "Trustworthy AI development guidelines for human system interaction," in *Proc. 13th Int. Conf. Hum. Syst. Interact. (HSI)*, Jun. 2020, pp. 130–136.

[17] S. T. Margulis, "On the status and contribution of Westin's and Altman's theories of privacy," *J. Soc. Issues*, vol. 59, no. 2, pp. 411–429, Jul. 2003, doi: 10.1111/1540-4560.00071.

[18] H. J. Smith, S. J. Milberg, and S. J. Burke, "Information privacy: Measuring individuals' concerns about organizational practices," *MIS Quart.*, vol. 20, no. 2, pp. 167–196, Jun. 1996, doi: 10.2307/249477.

[19] E. F. Stone, H. G. Gueutal, D. G. Gardner, and S. McClure, "A field experiment comparing information-privacy values, beliefs, and attitudes across several types of organizations," *J. Appl. Psychol.*, vol. 68, no. 3, pp. 459–468, Aug. 1983, doi: 10.1037/0021-9010.68.3.459.

[20] K. Beckers, *Pattern and Security Requirements: Engineering-Based Establishment of Security Standards*. Cham, Switzerland: Springer, 2015. [Online]. Available: https://books.google.ca/books?id=DvdICAAAQBAJ

[21] D. Proske, "Categorization of safety and risk," in *Perceived Safety: A Multidisciplinary Perspective* (Risk Engineering), M. Raue, B. Streicher, and E. Lermer, Eds. Cham, Switzerland: Springer, 2019, pp. 15–26, doi: 10.1007/978-3-030-11456-5_2.

[22] C. Stachl, Q. Au, R. Schoedel, S. D. Gosling, G. M. Harari, D. Buschek, S. T. Völkel, T. Schuwerk, M. Oldemeier, T. Ullmann, H. Hussmann, B. Bischl, and M. Bühner, "Predicting personality from patterns of behavior collected with smartphones," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 30, pp. 17680–17687, Jul. 2020.

[23] M. Altman, A. Wood, D. R. O'Brien, and U. Gasser, "Practical approaches to big data privacy over time," *Int. Data Privacy Law*, vol. 8, no. 1, pp. 29–51, Feb. 2018.

[24] J. Jia and N. Z. Gong, "AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning," in *Proc. 27th USENIX Secur. Symp.*, Aug. 2018, pp. 513–529, Accessed: Jul. 13, 2022. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/jia-jinyuan

[25] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18, doi: 10.1109/SP.2017.41.

[26] "Resolution on privacy by design," in *Proc. 32nd Int. Conf. Data Protection Privacy Commissioners*, Oct. 2010. [Online]. Available: http://globalprivacyassembly.org/wpcontent/uploads/2015/02/32-Conference-Israel-resolution-on-Privacy-by-Design.pdf

[27] *Regulation on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text With EEA Relevance)*. Accessed: Jun. 30, 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[28] M. Phillips, "International data-sharing norms: From the OECD to the general data protection regulation (GDPR)," *Hum. Genet.*, vol. 137, no. 8, pp. 575–582, Aug. 2018.
[29] E. Goldman, "An introduction to the California consumer privacy act (CCPA)," School Law, Santa Clara Univ., Santa Clara, CA, USA, Tech. Rep., 2020.
[30] The Office of the Privacy Commissioner of Canada. (Sep. 16, 2011). *PIPEDA Fair Information Principles*. Accessed: Jul. 22, 2022. [Online]. Available: https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/
[31] D. Jaar and P. E. Zeller, "Canadian privacy law: The personal information protection and electronic documents act (PIPEDA)," *Int.-House Counsel J.*, vol. 2, no. 7, pp. 1135–1146, 2009.
[32] S. D. C. D. Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data privacy: Definitions and techniques," *Int. J. Uncertain Fuzziness Knowl. Based Syst.*, vol. 20, no. 6, pp. 793–818, 2012.
[33] E. McCallister, T. Grance, and K. A. Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information*, vol. 800, no. 122. Collingdale, PA, USA: Diane Publishing, 2010.
[34] T. Dalenius, "Finding a needle in a haystack or identifying anonymous census records," *J. Off. Statist.*, vol. 2, no. 3, p. 329, Sep. 1986.
[35] L. Sweeney, "Simple demographics often identify people uniquely," *Health*, vol. 671, no. 2000, pp. 1–34, 2000.
[36] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, no. 1, p. 1736, Mar. 2013, doi: 10.1038/srep01376.
[37] P. S. Chauhan and N. Kshetri, "2021 state of the practice in data privacy and security," *Computer*, vol. 54, no. 8, pp. 125–132, Aug. 2021, doi: 10.1109/MC.2021.3083916.
[38] M. B. Forcier, H. Gallois, S. Mullan, and Y. Joly, "Integrating artificial intelligence into health care through data access: Can the GDPR act as a beacon for policymakers?" *J. Law Biosci.*, vol. 6, no. 1, pp. 317–335, Oct. 2019.
[39] (Oct. 17, 2022). *What are 'Controllers' and 'Processors'?* Accessed: Oct. 20, 2022. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/controllers-and-processors/what-are-controllers-and-processors/
[40] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *J. Data Warehousing*, vol. 5, no. 4, pp. 13–22, 2000.
[41] M. Haakman, L. Cruz, H. Huijgens, and A. van Deursen, "AI lifecycle models need to be revised. An exploratory study in fintech," 2020, *arXiv:2010.02716*.
[42] R. Schwartz, L. Down, A. Jonas, and E. Tabassi, "A proposal for identifying and managing bias within artificial intelligence," Inf. Technol. Lab., Nat. Inst. Standards Technol., Gaithersburg, MD, USA, 2021, pp. 1–24.
[43] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques* (Database Management Systems), vol. 5, no. 4, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2011, pp. 83–124.
[44] J. Tsay, A. Braz, M. Hirzel, A. Shinnar, and T. Mummert, "AIMMX: Artificial intelligence model metadata extractor," in *Proc. 17th Int. Conf. Mining Softw. Repositories*, 2020, pp. 81–92.
[45] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. S. Mittal, and V. Munigala, "Overview and importance of data quality for machine learning tasks," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery, Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 3561–3562, doi: 10.1145/3394486.3406477.
[46] S. M. H. Fard, A. Hamzeh, and S. Hashemi, "Using reinforcement learning to find an optimal set of features," *Comput. Math. Appl.*, vol. 66, no. 10, pp. 1892–1904, Dec. 2013.
[47] B. Blobel, K. Engel, and P. Pharowe, "Semantic interoperability," *Methods Inf. Med.*, vol. 45, no. 4, pp. 343–353, 2006.
[48] M. L. Zeng, "Interoperability," *Knowl. Org.*, vol. 46, no. 2, pp. 122–146, 2019.
[49] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.
[50] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE 31st Comput. Secur. Found. Symp. (CSF)*, Jul. 2018, pp. 268–282.
[51] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 33–44.
[52] D. Slack, S. A. Friedler, C. Scheidegger, and C. D. Roy, "Assessing the local interpretability of machine learning models," 2019, *arXiv:1902.03501*.
[53] F. Hoffmann, T. Bertram, R. Mikut, M. Reischl, and O. Nelles, "Benchmarking in classification and regression," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 5, Sep. 2019, Art. no. e1318.
[54] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *Int. J. Comput. Commun.*, vol. 5, no. 1, pp. 27–34, 2011.
[55] L. Baier, N. Kühl, and G. Satzger, "How to cope with change? Preserving validity of predictive services over time," in *Proc. 52nd Annu. Hawaii Int. Conf. Syst. Sci.*, 2019, pp. 1085–1094.
[56] W. Hummer, V. Muthusamy, T. Rausch, P. Dube, K. El Maghraoui, A. Murthi, and P. Oum, "ModelOps: Cloud-based lifecycle management for reliable and trusted AI," in *Proc. IEEE Int. Conf. Cloud Eng. (IC2E)*, Jun. 2019, pp. 113–120.
[57] S. J. De and D. Le Métayer, "PRIAM: A privacy risk analysis methodology," in *Data Privacy Management and Security Assurance*. Cham, Switzerland: Springer, Sep. 2016, pp. 221–229.
[58] S. M. H. Fard, H. Karimimpour, A. Dehghantanha, A. N. Jahromi, and G. Srivastava, "Ensemble sparse representation-based cyber threat hunting for security of smart cities," *Comput. Electr. Eng.*, vol. 88, Dec. 2020, Art. no. 106825.
[59] D. Liginlal, I. Sim, and L. Khansa, "How significant is human error as a cause of privacy breaches? An empirical study and a framework for error management," *Comput. Secur.*, vol. 28, nos. 3–4, pp. 215–228, May 2009, doi: 10.1016/j.cose.2008.11.003.
[60] W. Tounsi and H. Rais, "A survey on technical threat intelligence in the age of sophisticated cyber attacks," *Comput. Secur.*, vol. 72, pp. 212–233, Jan. 2018.
[61] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! A survey of attacks on private data," *Annu. Rev. Statist. Appl.*, vol. 4, no. 1, pp. 61–84, Mar. 2017.
[62] D. Irani, S. Webb, K. Li, and C. Pu, "Modeling unintended personal-information leakage from multiple online social networks," *IEEE Internet Comput.*, vol. 15, no. 3, pp. 13–19, May 2011, doi: 10.1109/MIC.2011.25.
[63] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, "A survey on data leakage prevention systems," *J. Netw. Comput. Appl.*, vol. 62, pp. 137–152, Feb. 2016, doi: 10.1016/j.jnca.2016.01.008.
[64] *Grand Theft Data—Data Exfiltration Study: Actors, Tactics, and Detection*, McAfee, San Jose, CA, USA, 2015.
[65] J. X. Jiang and G. Bai, "Evaluation of causes of protected health information breaches," *J. Amer. Med. Assoc. Int. Med.*, vol. 179, no. 2, pp. 265–267, Feb. 2019, doi: 10.1001/jamainternmed.2018.5295.
[66] M. J. Culnan and C. C. Williams, "How ethics can enhance organizational privacy: Lessons from the choicepoint and TJX data breaches," *MIS Quart.*, vol. 33, no. 4, pp. 673–687, Dec. 2009, doi: 10.2307/20650322.
[67] M. M. Merener, "Theoretical results on de-anonymization via linkage attacks," *Trans. Data Privacy*, vol. 5, no. 2, pp. 377–402, Aug. 2012.
[68] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 265–273.
[69] J. Henriksen-Bulmer and S. Jeary, "Re-identification attacks—A systematic literature review," *Int. J. Inf. Manage.*, vol. 36, no. 6, pp. 1184–1192, Dec. 2016, doi: 10.1016/j.ijinfomgt.2016.08.002.
[70] A. A. Hussien, N. Hamza, and H. A. Hefny, "Attacks on anonymization-based privacy-preserving: A survey for data mining and data publishing," *J. Inf. Secur.*, vol. 4, no. 2, pp. 101–112, Apr. 2013, doi: 10.4236/jis.2013.42012.
[71] M. Maouche, S. B. Mokhtar, and S. Bouchenak, "AP-Attack: A novel user re-identification attack on mobility datasets," in *Proc. 14th EAI Int. Conf. Mobile Ubiquitous Syst., Comput., Netw. Services (MobiQuitous)*. New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 48–57, doi: 10.1145/3144457.3144494.

[72] S. Gambs, A. Gmati, and M. Hurfin, "Reconstruction attack through classifier analysis," in *Proc. IFIP Annu. Conf. Data Appl. Secur. Privacy*, 2012, pp. 274–281.

[73] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst. (PODS)*. New York, NY, USA: Association for Computing Machinery, Jun. 2003, pp. 202–210, doi: 10.1145/773153.773173.

[74] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and privacy in machine learning," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Apr. 2018, pp. 399–414, doi: 10.1109/EuroSP.2018.00035.

[75] M. Veale, R. Binns, and L. Edwards, "Algorithms that remember: Model inversion attacks and data protection law," *Phil. Trans. Roy. Soc. A, Math. Phys. Eng. Sci.*, vol. 376, no. 2133, Nov. 2018, Art. no. 20180083.

[76] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: Association for Computing Machinery, Oct. 2015, pp. 1322–1333, doi: 10.1145/2810103.2813677.

[77] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "MemGuard: Defending against black-box membership inference attacks via adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 259–274.

[78] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998, doi: 10.1023/A:1009715923555.

[79] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[80] Y. K. Dwivedi et al., "Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int. J. Inf. Manage.*, vol. 57, Apr. 2021, Art. no. 101994, doi: 10.1016/j.ijinfomgt.2019.08.002.

[81] A. Elmes et al., "Accounting for training data error in machine learning applied to Earth observations," *Remote Sens.*, vol. 12, no. 6, p. 1034, Jan. 2020, doi: 10.3390/rs12061034.

[82] K. M. Kostick-Quenet, I. G. Cohen, S. Gerke, B. Lo, J. Antaki, F. Movahedi, H. Njah, L. Schoen, J. E. Estep, and J. S. Blumenthal-Barby, "Mitigating racial bias in machine learning," *J. Law, Med., Ethics*, vol. 50, no. 1, pp. 92–100, 2022, doi: 10.1017/jme.2022.13.

[83] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput. Math. Org. Theory*, vol. 25, no. 3, pp. 319–335, Sep. 2019, doi: 10.1007/s10588-018-9266-8.

[84] B. Bhushan, G. Sahoo, and A. K. Rai, "Man-in-the-middle attack in wireless and computer networking—A review," in *Proc. 3rd Int. Conf. Adv. Comput., Commun. Autom. (ICACCA)*, Sep. 2017, pp. 1–6.

[85] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2018, pp. 19–35, doi: 10.1109/SP.2018.00057.

[86] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE J. Biomed. Health Inf.*, vol. 19, no. 6, pp. 1893–1905, Nov. 2015, doi: 10.1109/JBHI.2014.2344095.

[87] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*.

[88] H. S. M. Lim and A. Taeihagh, "Algorithmic decision-making in AVs: Understanding ethical and technical concerns for smart cities," *Sustainability*, vol. 11, no. 20, p. 5791, Oct. 2019.

[89] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Commun. ACM*, vol. 61, no. 7, pp. 56–66, Jun. 2018.

[90] B. Biggio, I. Corona, B. Nelson, B. I. P. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli, "Security evaluation of support vector machines in adversarial environments," in *Support Vector Machines Applications*. Cham, Switzerland: Springer, Jan. 2014, pp. 105–153.

[91] B. Nelson, B. I. P. Rubinstein, L. Huang, A. D. Joseph, S. J. Lee, S. Rao, and J. D. Tygar, "Query strategies for evading convex-inducing classifiers," *J. Mach. Learn. Res.*, vol. 13, no. 5, pp. 1293–1332, 2012.

[92] R. N. Rodrigues, L. L. Ling, and V. Govindaraju, "Robustness of multimodal biometric fusion methods against spoof attacks," *J. Vis. Lang., Comput.*, vol. 20, no. 3, pp. 169–179, Jun. 2009.

[93] (Jan. 4, 2021). *What is Automated Individual Decision-Making and Profiling?* Accessed: Jul. 26, 2022. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-is-automated-individual-decision-making-and-profiling/

[94] K. Yeung, "Recommendation of the council on artificial intelligence (OECD)," *Int. Legal Mater.*, vol. 59, no. 1, pp. 27–34, Feb. 2020, doi: 10.1017/ilm.2020.5.

[95] C. Molnar, *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. 2020. [Online]. Available: https://Lulu.com

[96] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[97] A. J. London, "Artificial intelligence and black-box medical decisions: Accuracy versus explainability," *Hastings Center Rep.*, vol. 49, no. 1, pp. 15–21, Jan. 2019, doi: 10.1002/hast.973.

[98] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Market. Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.

[99] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *Int. Data Privacy Law*, vol. 7, no. 2, pp. 76–99, May 2017.

[100] A. Selbst and J. Powles, "'Meaningful information' and the right to explanation," in *Proc. 1st Conf. Fairness, Accountability Transparency (PMLR)*, Jan. 2018, p. 48, Accessed: Jul. 26, 2022. [Online]. Available: https://proceedings.mlr.press/v81/selbst18a.html

[101] B. Wagner, "Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications," Council European, Strasbourg, France, Tech. Rep., 2016.

[102] *Government Bill (House of Commons) C-27 (44–1)-First Reading-Digital Charter Implementation Act, 2022-Parliament of Canada*. Accessed: Jul. 5, 2022. [Online]. Available: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading

[103] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Society (AIES)*. New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 231–241, doi: 10.1145/3461702.3462533.

[104] Y. Wang, H. Qian, and C. Miao, "DualCF: Efficient model extraction attack from counterfactual explanations," in *Proc. ACM Conf. Fairness, Accountability, Transparency*. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 1318–1329, doi: 10.1145/3531146.3533188.

[105] A. Dabrowski, G. Merzdovnik, J. Ullrich, G. Sendera, and E. Weippl, "Measuring cookies and web privacy in a post-GDPR world," in *Proc. Int. Conf. Passive Act. Netw. Meas.*, Mar. 2019, pp. 258–270.

[106] P. De Hert, V. Papakonstantinou, G. Malgieri, L. Beslay, and I. Sanchez, "The right to data portability in the GDPR: Towards user-centric interoperability of digital services," *Comput. Law, Secur. Rev.*, vol. 34, no. 2, pp. 193–203, Apr. 2018.

[107] V. B. Livshits and M. S. Lam, "Finding security vulnerabilities in Java applications with static analysis," in *Proc. USENIX Secur. Symp.*, Jul. 2005, p. 18.

[108] A. Mallik, "Man-in-the-middle-attack: Understanding in simple words," *J. Pendidikan Teknol. Inf.*, vol. 2, no. 2, pp. 109–134, 2019.

[109] *OECD Legal Instruments*. Accessed: Jul. 22, 2022. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[110] D. Mishra, A. Mishra, and A. Yazici, "Successful requirement elicitation by combining requirement engineering techniques," in *Proc. 1st Int. Conf. Appl. Digit. Inf. Web Technol. (ICADIWT)*, Aug. 2008, pp. 258–263, doi: 10.1109/ICADIWT.2008.4664355.

[111] I. Johnston and P. Sawyer, *Requirements Engineering: A Good Practice Guide*. Beijing, China: China Machine Press, May 1997, pp. 17–206.

[112] Department of Homeland Security. (Aug. 2015). *Privacy Impact Assessment Guidance*. Accessed: Jul. 27, 2022. [Online]. Available: https://www.dhs.gov/publication/privacy-impact-assessment-guidance

[113] M. Elkhodr, B. Alsinglawi, and M. Alshehri, "A privacy risk assessment for the Internet of Things in healthcare," in *Applications of Intelligent Technologies in Healthcare*. Cham, Switzerland: Springer, Nov. 2018, pp. 47–54.
[114] K. Vemou and M. Karyda, "Evaluating privacy impact assessment methods: Guidelines and best practice," *Inf. Comput. Secur.*, vol. 28, no. 1, pp. 35–53, 2019.
[115] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–38, May 2019.
[116] (Feb. 11, 2021). *Principle (C): Data Minimisation*. Accessed: Aug. 1, 2022. [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/
[117] A. Goldsteen, G. Ezov, R. Shmelkin, M. Moffie, and A. Farkash, "Data minimization for GDPR compliance in machine learning models," *AI Ethics*, vol. 2, pp. 477–491, Sep. 2021, doi: 10.1007/s43681-021-00095-8.
[118] C. Castelluccia, M. Cunche, D. Le Metayer, and V. Morel, "Enhancing transparency and consent in the IoT," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroSPW)*, Apr. 2018, pp. 116–119.
[119] E. Costante, Y. Sun, M. Petković, and J. D. Hartog, "A machine learning solution to assess privacy policy completeness: (Short paper)," in *Proc. ACM Workshop Privacy Electron. Soc. (WPES)*. New York, NY, USA: Association for Computing Machinery, Oct. 2012, pp. 91–96, doi: 10.1145/2381966.2381979.
[120] A. Sokolovska and L. Kocarev, "Integrating technical and legal concepts of privacy," *IEEE Access*, vol. 6, pp. 26543–26557, 2018.
[121] A. Siddiqa, A. Karim, and A. Gani, "Big data storage technologies: A survey," *Front. Inf. Technol. Electron. Eng.*, vol. 18, no. 8, pp. 1040–1070, 2017.
[122] H. J. Pandit, C. Debruyne, D. O'Sullivan, and D. Lewis, "An exploration of data interoperability for GDPR," *Int. J. Standardization Res.*, vol. 16, no. 1, pp. 1–21, Jan. 2018.
[123] A. Jaleel, T. Mahmood, M. A. Hassan, G. Bano, and S. K. Khurshid, "Towards medical data interoperability through collaboration of healthcare devices," *IEEE Access*, vol. 8, pp. 132302–132319, 2020.
[124] E. J. A. Folmer and J. Verhoosel, "State of the art on semantic IS standardization, interoperability & quality," TNO, Universiteit Twente, NOiV, CTIT, Enschede, The Netherlands, Tech. Rep., 2011.
[125] L. Bezuidenhout, "Being fair about the design of FAIR data standards," *Digit. Government Res. Pract.*, vol. 1, no. 3, pp. 18.1–18.7, Sep. 2020, doi: 10.1145/3399632.
[126] N. Gruschka, V. Mavroeidis, K. Vishi, and M. Jensen, "Privacy issues and data protection in big data: A case study analysis under GDPR," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 5027–5033.
[127] P. Jain, M. Gyanchandani, and N. Khare, "Big data privacy: A technological perspective and review," *J. Big Data*, vol. 3, no. 1, p. 25, Nov. 2016, doi: 10.1186/s40537-016-0059-y.
[128] F. Chen, B. Bruhadeshwar, and A. X. Liu, "Cross-domain privacy-preserving cooperative firewall optimization," *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 857–868, Jun. 2013.
[129] N. Kasrin, M. Qureshi, S. Steuer, and D. Nicklas, "Semantic data management for experimental manufacturing technologies," *Datenbank-Spektrum*, vol. 18, no. 1, pp. 27–37, Mar. 2018, doi: 10.1007/s13222-018-0274-0.
[130] E. F. Silva, D. C. Muchaluat-Saade, and N. C. Fernandes, "ACROSS: A generic framework for attribute-based access control with distributed policies for virtual organizations," *Future Gener. Comput. Syst.*, vol. 78, pp. 1–17, Jan. 2018, doi: 10.1016/j.future.2017.07.049.
[131] T. Wang, J. Zhao, Z. Hu, X. Yang, X. Ren, and K.-Y. Lam, "Local differential privacy for data collection and analysis," *Neurocomputing*, vol. 426, pp. 114–133, Feb. 2021, doi: 10.1016/j.neucom.2020.09.073.
[132] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney, "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM J. Res. Develop.*, vol. 63, nos. 4–5, pp. 1–6, Jul. 2019.
[133] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *Int. J. Adv. Softw.*, vol. 10, no. 1, pp. 1–20, 2017.
[134] G. Jagannathan and R. N. Wright, "Privacy-preserving imputation of missing data," *Data Knowl. Eng.*, vol. 65, no. 1, pp. 40–56, Apr. 2008, doi: 10.1016/j.datak.2007.06.013.
[135] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, vol. 793. Hoboken, NJ, USA: Wiley, 2019.
[136] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.
[137] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM Trans. Knowl. Discovery Data*, vol. 10, no. 1, pp. 1–51, Jul. 2015.
[138] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009.
[139] H. Zenati, M. Romain, C. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially learned anomaly detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 727–736.
[140] S. Shahriar, A. R. Al-Ali, A. H. Osman, S. Dhou, and M. Nijim, "Prediction of EV charging behavior using machine learning," *IEEE Access*, vol. 9, pp. 111576–111586, 2021, doi: 10.1109/ACCESS.2021.3103119.
[141] T. Neubauer and J. Heurix, "A methodology for the pseudonymization of medical data," *Int. J. Med. Inf.*, vol. 80, no. 3, pp. 190–204, Mar. 2011, doi: 10.1016/j.ijmedinf.2010.10.016.
[142] D. N. Jaidan, M. Carrere, Z. Chemli, and R. Poisvert, "Data anonymization for privacy aware machine learning," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.*, Jan. 2019, pp. 725–737.
[143] J. Soria-Comas and J. Domingo-Ferrer, "Mitigating the curse of dimensionality in data anonymization," in *Proc. Int. Conf. Modeling Decis. Artif. Intell.*, vol. 2019, pp. 346–355.
[144] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, Apr. 2016, doi: 10.1016/j.neucom.2015.08.104.
[145] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?" in *Proc. 3rd ACM Workshop Cloud Comput. Secur. Workshop*, Oct. 2011, pp. 113–124.
[146] P. Bashivan et al., "Adversarial feature desensitization," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2021, pp. 10665–10677, Accessed: Oct. 21, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/587b7b833034299fdd5f4b10e7dc9fca-Abstract.html
[147] F. M. Zanzotto, "Viewpoint: Human-in-the-loop artificial intelligence," *J. Artif. Intell. Res.*, vol. 64, pp. 243–252, Feb. 2019.
[148] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, and A. Parameswaran, "Accelerating human-in-the-loop machine learning: Challenges and opportunities," in *Proc. 2nd Workshop Data Manage. End-To-End Mach. Learn.*, 2018, pp. 1–4.
[149] E. Mortaz, "Imbalance accuracy metric for model selection in multi-class imbalance classification problems," *Knowl.-Based Syst.*, vol. 210, Dec. 2020, Art. no. 106490, doi: 10.1016/j.knosys.2020.106490.
[150] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Statist. Comput.*, vol. 21, no. 2, pp. 137–146, Apr. 2011, doi: 10.1007/s11222-009-9153-8.
[151] R. G. Mantovani, T. Horváth, R. Cerri, S. B. Junior, J. Vanschoren, and A. C. P. D. L. F. D. Carvalho, "An empirical study on hyperparameter tuning of decision trees," 2019, *arXiv:1812.02207*.
[152] H. Alibrahim and S. A. Ludwig, "Hyperparameter optimization: Comparing genetic algorithm against grid search and Bayesian optimization," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2021, pp. 1551–1559, doi: 10.1109/CEC45853.2021.9504761.
[153] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, Jan. 2004, doi: 10.1021/ci0342472.
[154] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[155] C. F. G. D. Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–25, Jan. 2022, doi: 10.1145/3510413.
[156] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019.
[157] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.
[158] G. Peake and J. Wang, "Explanation mining: Post hoc interpretability of latent factor models for recommendation systems," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2060–2069.

[159] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[160] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 4768–4777.

[161] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, *arXiv:1802.03888*.

[162] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017.

[163] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[164] X. Jiang and Z. Ge, "Data augmentation classifier for imbalanced fault classification," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1206–1217, Jul. 2021, doi: 10.1109/TASE.2020.2998467.

[165] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning From Imbalanced Data Sets*. Cham, Switzerland: Springer, Accessed: Aug. 3, 2022. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-98074-4

[166] I. Gat, I. Schwartz, A. Schwing, and T. Hazan, "Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, Dec. 2020, pp. 3197–3208. Accessed: Aug. 3, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/20d749bc05f47d2bd3026ce457dcfd8e-Abstract.html

[167] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Secur. Privacy*, vol. 17, no. 2, pp. 49–58, Mar. 2019.

[168] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb, "Identification using encrypted biometrics," in *Proc. Int. Conf. Comput. Anal. Images Pattern*. Berlin, Germany: Springer, Aug. 2013, pp. 440–448.

[169] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, Aug. 2014, doi: 10.1561/0400000042.

[170] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2000, pp. 439–450.

[171] J. Kim and W. Winkler, "Multiplicative noise for masking continuous data," *Statistics*, vol. 1, no. 9, pp. 1–18, Apr. 2003.

[172] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 475–489.

[173] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, Apr. 2017, pp. 1273–1282. Accessed: Aug. 3, 2022. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[174] G. Zyskind, O. Nathan, and A. Pentland, "Enigma: Decentralized computation platform with guaranteed privacy," 2015, *arXiv:1506.03471*.

[175] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[176] X. Meng, "Scalable simple random sampling and stratified sampling," in *Proc. 30th Int. Conf. Mach. Learn.*, May 2013, pp. 531–539. Accessed: Nov. 19, 2022. [Online]. Available: https://proceedings.mlr.press/v28/meng13a.html

[177] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *J. Exp., Theor. Artif. Intell.*, vol. 12, no. 1, pp. 1–12, Jan. 2000, doi: 10.1080/095281300146272.

[178] A. B. Haque, A. K. M. N. Islam, and P. Mikalef, "Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research," *Technol. Forecast. Soc. Change*, vol. 186, Jan. 2023, Art. no. 122120, doi: 10.1016/j.techfore.2022.122120.

[179] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017.

[180] M. Conti, T. Dargahi, and A. Dehghantanha, "Cyber threat intelligence: Challenges and opportunities," in *Cyber Threat Intelligence* (Advances in Information Security), A. Dehghantanha, M. Conti, and T. Dargahi, Eds. Cham, Switzerland: Springer, Apr. 2018, pp. 1–6, doi: 10.1007/978-3-319-73951-9_1.

[181] G. Moallem, D. D. Pathirage, J. Reznick, J. Gallagher, and H. Sari-Sarraf, "An explainable deep vision system for animal classification and detection in trail-camera images with automatic post-deployment retraining," *Knowl.-Based Syst.*, vol. 216, Mar. 2021, Art. no. 106815, doi: 10.1016/j.knosys.2021.106815.

[182] C. K. N. C. K. Mohd and F. Shahbodin, "Personalized learning environment: Alpha testing, beta testing & user acceptance test," *Proc., Soc. Behav. Sci.*, vol. 195, pp. 837–843, Jul. 2015, doi: 10.1016/j.sbspro.2015.06.319.

[183] *DP Notice Generator*. Accessed Nov. 19, 2022. [Online]. Available: https://apps.pdpc.gov.sg/dp-notice-generator

[184] D. A. Audich, R. Dara, and B. Nonnecke, "Improving readability of online privacy policies through DOOP: A domain ontology for online privacy," *Digital*, vol. 1, no. 4, pp. 198–215, Nov. 2021, doi: 10.3390/digital1040015.

[185] M. Cunneen, M. Mullins, and F. Murphy, "Autonomous vehicles and embedded artificial intelligence: The challenges of framing machine driving decisions," *Appl. Artif. Intell.*, vol. 33, no. 8, pp. 706–731, Jul. 2019, doi: 10.1080/08839514.2019.1600301.

[186] H. Li, L. Yu, and W. He, "The impact of GDPR on global technology development," *J. Global Inf. Technol. Manage.*, vol. 22, no. 1, pp. 1–6, Jan. 2019, doi: 10.1080/1097198X.2019.1569186.

[187] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, 2020, doi: 10.1109/TIFS.2020.2988575.

[188] H. Kim, S. Kim, J. Y. Hwang, and C. Seo, "Efficient privacy-preserving machine learning for blockchain network," *IEEE Access*, vol. 7, pp. 136481–136495, 2019, doi: 10.1109/ACCESS.2019.2940052.

[189] O. Fadi, Z. Karim, E. G. Abdellatif, and B. Mohammed, "A survey on blockchain and artificial intelligence technologies for enhancing security and privacy in smart environments," *IEEE Access*, vol. 10, pp. 93168–93186, 2022, doi: 10.1109/ACCESS.2022.3203568.

[190] R. Bosri, M. S. Rahman, M. Z. A. Bhuiyan, and A. Al Omar, "Integrating blockchain with artificial intelligence for privacy-preserving recommender systems," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1009–1018, Apr. 2021, doi: 10.1109/TNSE.2020.3031179.

[191] G. P. Freund, P. B. Fagundes, and D. D. J. de Macedo, "An analysis of blockchain and GDPR under the data lifecycle perspective," *Mobile Netw. Appl.*, vol. 26, no. 1, pp. 266–276, Feb. 2021, doi: 10.1007/s11036-020-01646-9.

[192] A. Bröring, S. Schmid, C. Schindhelm, A. Khelil, S. Käbisch, D. Kramer, D. Le Phuoc, J. Mitic, D. Anicic, and E. Teniente, "Enabling IoT ecosystems through platform interoperability," *IEEE Softw.*, vol. 34, no. 1, pp. 54–61, Jan. 2017, doi: 10.1109/MS.2017.2.

SAKIB SHAHRIAR received the B.S. and M.S. degrees in computer engineering from the American University of Sharjah, United Arab Emirates, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree with the Data Management and Privacy Governance Laboratory, School of Computer Science, University of Guelph. His current research interests include machine learning, natural language processing, information privacy, and deep learning.

SONAL ALLANA received the bachelor's degree in computer engineering from the University of Mumbai and the master's degree in computing from the National University of Singapore. She is currently pursuing the Ph.D. degree with the Data Management and Privacy Governance Laboratory, School of Computer Science, University of Guelph. Her current research interests include trustworthiness and fairness in AI systems, privacy, and cybersecurity.

SEYED MEHDI HAZRATIFARD received the M.Sc. and Ph.D. degrees in artificial intelligence from Shiraz University. He was a Computer Scientist at various companies, such as Irancell and Soshianest, which helped him to build a strong background in real applications of machine learning. He was a Postdoctoral Researcher with the University of Guelph. Currently, he is a Postdoctoral Researcher with the University of Victoria. His current research interests include machine learning applications, such as providing security and privacy in smart environments and using machine learning in forecasting and classification.

ROZITA DARA is currently the Director of the Data Management and Privacy Governance Research Program, University of Guelph. Since joining the University of Guelph, she has been spearheading several initiatives in the area of information privacy and the Internet of Things, including data platforms, trust management systems, and data and technology trust frameworks. She has also led and contributed to several national and global efforts on the standardization of information privacy, the Internet of Things, digital platforms (e.g., blockchain), and trustable artificial intelligence. Her current research interests include information privacy, intelligent systems, and data governance. She has served as a guest editor for several special issues of IEEE and Elsevier journals and IEEE conference chair.