Privacy and Artificial Intelligence

James Curzon, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib

***Abstract*—Artificial intelligence is a rapidly developing field of research with many practical applications. Congruent to advances in technologies that enable big data, deep learning, and neural networks to train, learn, and predict, artificial intelligence creates new risks that are difficult to predict and manage. Such risks include economic turmoil, existential crises, and the dissolution of individual privacy. If unchecked, the capabilities of artificially intelligent systems could pose a fundamental threat to privacy in their operation or these systems may leak information under adversarial conditions. In this article, we survey the literature and provide various scenarios for the use of artificial intelligence, highlighting potential risks to privacy and offering various mitigating strategies. For the purpose of this research, a North American perspective of privacy is adopted.**

***Impact statement*—While an appreciation of the privacy risks associated with artificial intelligence is important, a thorough understanding of the assortment of different technologies that comprise artificial intelligence better prepares those implementing such systems in assessing privacy impacts. This can be achieved through the independent consideration of each constituent of an artificially intelligent system and its interactions. Under individual consideration, privacy-enhancing tools can be applied in a targeted manner to reduce the risk associated with specific components of an artificially intelligent system. A generalized North American approach to assess privacy risks in such systems is proposed that will retain applicability as the field of research evolves and can be adapted to account for various sociopolitical influences. With such an approach, privacy risks in artificial intelligent systems can be well understood, measured, and reduced.**

***Index Terms*—Artificial intelligence (AI), knowledge representation, machine learning, natural language processing (NLP), privacy.**

I. Introduction

ARTIFICIAL intelligence (AI) is one of several modern approaches to achieving human-equivalent machine intelligence [1]. Other approaches are brain emulation, biological cognition, and human–machine interfaces through the organization of networks (e.g., the Internet) into a form of collective intelligence. Of these technologies, AI has experienced significant successes and has a dominant presence in daily life.

Manuscript received September 8, 2020; revised February 2, 2021 and April 6, 2021; accepted June 1, 2021. Date of publication June 9, 2021; date of current version August 20, 2021. This work was supported by the Office of the Privacy Commissioner of Canada. This article was recommended for publication by Associate Editor C. Wagner upon evaluation of the reviewers' comments. *(Corresponding author: Khalil El-Khatib.)*

James Curzon is with the University of Ontario Institute of Technology, Oshawa, ON L1G 0C5, Canada (e-mail: james.curzon@ontariotechu.net).

Tracy Ann Kosa is with the Seattle University School of Law, Seattle, WA 98122 USA (e-mail: kosat@seattleu.edu).

Rajen Akalu and Khalil El-Khatib are with the Faculty of Business and Information Technology, University of Ontario Institute of Technology, Oshawa, ON L1G 0C5, Canada (e-mail: rajen.akalu@uoit.ca; khalil.el-khatib@uoit.ca).

Digital Object Identifier 10.1109/TAI.2021.3088084

AI is described as a field of study focused on creating intelligent entities, with numerous applications in various domains, including playing games (e.g., chess, checkers, go), proving mathematical theorems, managing homes, and driving cars. At a high level, there are four approaches to AI when designing such a system [2]. The first approach is creating a system that thinks humanly, including activities in problem solving and learning. Second, a system can be designed to think rationally, focusing on perception, reasoning, and action. Third, a system can be designed to act humanly, attempting to mimic human behavior in machines (e.g., the Turing test). Finally, systems can be designed to act rationally, attempting to make rational decisions in uncertain environments.

Machine learning is a tool to help create and implement AI systems and leans heavily into statistical methods to accomplish its goals. While an AI system may perceive its environment with sensors and take actions with actuators, machine learning enables the system to learn from experiences. Specifically, research in the field of machine learning contributes to answers for important problems in AI [3], such as how systems can autonomously improve and what statistical-computational information-theoretic laws govern learning systems. A machine learning system is created through a training process to generate a model that predicts future outputs based on new inputs. Typically, machine learning models are created through supervised, unsupervised, semisupervised, or reinforcement learning. Recent trends in machine learning include increasing complex models running with access to increasing amounts of data [3].

While there are perceived benefits of developing human-equivalent machine intelligence, such as fostering the rapid development of human technological advancement, there are also a number of public concerns ranging from economic instability to apocalyptic consequences. This article will focus on a more direct and immediate consequence of smarter machines: a degradation to the privacy of both willing and unwilling participants of AI systems that operate on personal data such as facial or voice recognition systems. Privacy, in this context, is considered from a North American perspective where legislation is less stringent and privacy may be considered more of a commodity than a human right [4]. Public concern for privacy infringements with AI-enabled devices has been increasing, especially with the advent of personal assistants such as Google Home and Amazon Alexa. Stories and articles of these devices spying on individuals and recording voice patterns are not favorable for public acceptance of AI [5]. A controversial article by Wang and Kosinski [6] showed how deep neural networks can be trained to classify an individual's sexual orientation based on facial features, with an accuracy of 91%. Such technology has

direct consequences related to privacy, providing a mechanism to reveal the sexual orientation of an individual who may otherwise not wish their orientation revealed. More seriously, in repressive regimes, sexual orientation extends to a matter of life and death, and developments in AI pose a significant increase in risk for those who are oppressed. Further, the data used in training and the characteristics used in practice have direct consequences on privacy. As such, this article will provide a review of privacy, explore known privacy risks associated with AI, identify popular use cases, and discuss their associated risks.

The rest of this article is be organized as follows. Section II provides a baseline for understanding privacy, reviewing common definitions and regulations. Section III describes topics in AI and reviews each component of AI in its current state. Privacy risks and threats that are resultant for each component of AI are described in Section IV. Mitigating efforts in protecting privacy are discussed in Section V. Finally, Section VI concludes the article.

II. Privacy

Before investigating how concepts of privacy relate to the contributing technologies of AI, we seek to set out a basis for the term. One commonly cited definition of privacy is the state in which a person is not observed or disturbed, or to be free from public attention [7]. Notions of observation, attention, and disturbance must be adjusted for digital identities that are difficult to define, control, or understand. In this section, for the purpose of setting a baseline, we narrow our focus to American definitions of informational privacy as distinct. This article will focus on the North American approach to privacy law which has sought a balance between the privacy protection of individuals and legitimate business interests. This is to be contrasted with EU legislative approaches to privacy protection which have tended to focus on privacy as a human right.

A. Definitions

There are two distinct approaches to privacy in the American context. First, legal models of privacy, such as research by Margulis [8] or Solove [9], provide a groundwork for approaching privacy concepts. Second, a social lens, such as the studies provided by Kwasny *et al.* [10] or Smith *et al.* [11], that explores privacy in the context of individual versus group expectations.

Discussion of privacy commonly refers to the work of Westin and Altman, which is thoroughly described by Margulis [8]. Margulis summarized Westin's theory of privacy as to how an individual temporarily restricts access to themselves from others, identifying multiple states of privacy ranging from solitude to reserve. Altman's properties of privacy, as described by Margulis, revolve around the ability of an individual to selectively control access to themselves. The bidirectionality of privacy refers to how privacy considerations must include both inputs from others to an individual and the output from that individual to others. Individual and group privacy, or units of privacy, refers to what is considered the subject of privacy consideration. Privacy concepts apply equally to individuals as well as groups.

Solove [9], a legal scholar, approached privacy as a taxonomy of activities that work to degrade privacy. In this way, Solove seeks less to identify what privacy is, choosing to focus on what happens when it is missing.

Kwasny *et al.* [10] postulated technological interactions requiring an increasing amount of personal information disclosures, such as with online banking or social media, critically require privacy to be considered in the design and implementation phase. Smith *et al.* [11] similarly researched public opinion and identified domains of privacy. These domains are collection, internal or external secondary use of information, of errors in the information, improper access, reduced judgment in decisions, and aggregation.

An important distinction made in the study by Smith *et al.* and hinted at by Kwansy *et al.* is the transition from more standard definitions of privacy, per Altman, Westin, and Solove, to a more modern approach represented as "informational privacy." As we move deeper into the information age, public concerns of privacy are increasing along with the availability of data to organizations (both private and public) that collect it. These details and nuances will be explored next.

B. Information Privacy

The origins of privacy focused on data points, which are combined to represent some information about a given person or group of people. Information privacy moves the focus to the collection, use, and disclosure of that information within a given system.

Belanger and Crossler reviewed the field of information privacy in [12]. They defined it as the ability of an individual to exert control over data about themselves. Advances in information technology, such as the introduction and adoption of machine learning, bring additional complexity of attempts to provide privacy in such systems. Belanger and Crossler reviewed theoretical, structural, and societal aspects of privacy research and proposed categorization based on sensitivity: identifying, quasi-identifying, confidential, and nonconfidential [13].

Dalenius [14] described how seemingly unrelated attributes can be combined into identifiers; from assigning values to creating and sorting a matrix that reveals identical attribute values and also individuals with unique sets of attributes. The more attributes collected, the higher the risk of general identifiability.

C. Privacy Paradox

Worth discussing when reviewing privacy are the social perspectives on privacy issues, as public opinions on privacy are a driving factor of regulation [15]. While individuals often claim privacy is important, research shows behavioral patterns contrary to this opinion, as reviewed by Kokolakis in [16]. Dubbed the "privacy paradox," this disconnect between privacy attitudes and behavior is evident in many instances. Data breaches on voluntary social networking sites often result in public outcry over the loss of privacy as in the cases of Facebook [17].

In [18], Acquisti *et al.* outlined a parallel phenomenon: users who perceive themselves as having control over their own private information will be less concerned with the risks associated with

sharing that information. Referred to as the control paradox, Acquisti *et al.* found from experiments that while providing more privacy controls should be expected to improve privacy, the presence of these controls tended to lead individuals into making riskier decisions with their information. A possible solution proposes the use of radical transparency by organizations adhering to various regulations.

III Artificial Intelligence

Stanford University, in their AI100 study (100-year study on AI) [19], made an argument that there can be no concrete definition of AI, as the field is dynamic and continuously evolving. Instead, as the authors suggested, it may be more appropriate to define AI as the field of the study comprised of whatever it is that AI researchers are currently doing to simulate intelligence in machines using human intelligence as a benchmark. However, Russell and Norvig [2] outlined four approaches to AI: thinking humanly, acting humanly, thinking rationally, and acting rationally.

Systems designed to *think humanly* follow a cognitive modeling approach, where the input/output mechanisms in the system attempt to emulate the same process exhibited in humans. This notion of *cognition* is defined by Vernon [20] as the capacity for self-reliance, problem solving, and adaptive action or, in other words, the ability to infer a future state of the environment based on past experiences and the anticipated outcome from an action, where the action is chosen to achieve some goal. Cognitive models face two important issues: what the model is intended to achieve and how it should do so. This is to say, the model must achieve some behavior through some mechanism and the problem is referred to as the ultimate-proximate distinction. For a desired outcome or behavior, there are potentially many different applicable mechanisms. A human, for example, could draw from past experiences in an unrelated environment to find solutions (mechanisms) that solve a problem (behavior) in a different environment.

On the other hand, systems that *act humanly* are designed to emulate human behavior. Famously, this is known as the Turing test, where the system passes if an interrogator is unable to determine whether they are interacting with a computer or another human. These systems must be proficient in several areas: natural language processing (NLP), knowledge representation, automated reasoning, and learning. Knowledge representation is a crucial problem in AI that better enables systems to behave humanly. The study in knowledge representation is vast and covers many different topics from classical logic (propositional, first-order, second-order) to constraint programming to Bayesian networks [21].

A system that *thinks rationally* is designed to adhere to the laws of thought. These laws include [22] the law of contradiction, the law of excluded middle, and the principle of identity. The goal of these systems is to consistently construct correct argument structures given a correct premise. In other words, these are systems designed to strictly use logic to solve problems, which leads to some difficulties. One such difficulty is in how to transform informal knowledge (e.g., derived from unstructured data) into a meaningful logical notation. Another difficulty is in how to prioritize making a decision on information when the amount of information can be overwhelming. This leads to computation-time restrictions for what information should be considered relevant, especially in situations with uncertainty.

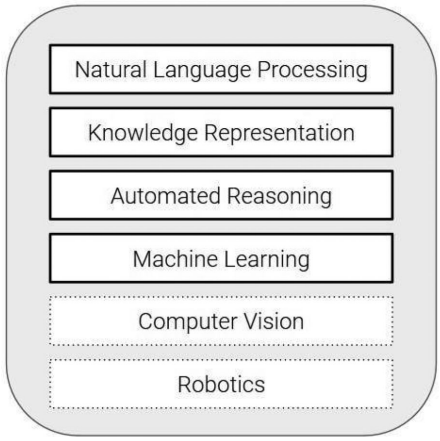

Fig. 1. Components of AI.

Finally, systems that *act rationally* are those that follow the rational agent approach. Such a system attempts to reach the best desirable outcome. This approach is similar to systems that are designed to think rationally, with the added feature of being able to operate and make the best decisions when faced with uncertainty. Designing systems to act rationally leads to a more general system following traditional sciences in behavior and thought.

Stone *et al.* [19] identified the major components of artificially intelligent systems include, at a minimum, NLP, knowledge representation, automated reasoning, and machine learning. Optional requirements include computer vision and robotics, and should the system be designed to pass the total Turing test [2] (see Fig. 1). Each of these components may pose its own privacy unique risks and considerations. We will briefly describe these components in this section, and explore their implications in the following section.

A. Computer Vision

Computer vision allows intelligent machines to perceive their environment and objects within their environment to make decisions in carrying out their objectives, such as a self-driving car slowing to a stop at a red light. The ability of computers to achieve the vision is not a challenge in simply designing artificial eyes, but rather in how images are processed to derive information. Improving computer vision is, then, the study of processing data in images to achieve object detection, motion tracking, or action recognition, among other goals [23].

B. Natural Language Processing

There is an enormous amount of knowledge stored as data in textual formats, and using various NLP techniques, all of which

can be explored to improve machine intelligence should machines be able to understand written language. The knowledge that is not saved textually can be translated into text through other means, such as speech recognition applications. An artificially intelligent system, then, would benefit immensely from being able to interpret natural languages, which is the goal of NLP.

C. Knowledge Representation

Knowledge representation translates real-world information into a machine-readable format. Two components of knowledge representation are inference and reasoning [24]. There are numerous techniques to represent knowledge for AI systems: lists, trees, semantic networks, schemas, rule-based, and logic-based [25]. Every technique, in some manner, uses a hierarchical structure, such as ontologies. Whatever method is chosen, artificially intelligent systems use some type of knowledge representation in order to learn, remember, and make decisions.

D. Automated Reasoning

The mechanism that allows an AI system to use the information to provide answers and draw conclusions is called automated reasoning [2]. Also referred to as automated deduction, applications for automated reasoning originated in proving mathematical theorems, as discussed by Bundy [26]. Bundy further described more modern capabilities of automated reasoning, which include proving nontrivial theorems, verifying system specification, synthesizing systems from specification, transforming a system into a more efficient equivalent, and, notably, common-sense reasoning in AI.

This area of research, of enabling systems to make their own deductions and reasoning, aligns with objectives of the field of machine common sense [27]. The current state of AI is hampered by the necessity to train on all possible scenarios in order to achieve robustness. Modern AI, dubbed "Narrow AI" or "Weak AI," has a limiting factor in a sort of "common-sense service," which would empower AI systems to adapt, reason, and communicate better. In fact, reliance on extensive supervised training is a problem extending to deep neural networks, as discussed by Trihn and Le [28]. This limitation prevents deep neural networks from being applied to unsupervised learning or areas with scarce labeled data. A modern test from Levesque [29], offered as an alternative Turing test, provides a benchmark for a system's reasoning capabilities against typed sentences in a series of multiple-choice questions.

E. Machine Learning

Machine learning, as the name suggests, is the study of how machines can improve through experience. Machine learning relies on prior knowledge, data representation, and feedback [2]. The previous topics in machine learning largely feed into machine learning, making it likely the most critical component of AI. Techniques used in machine learning are identical to those in the closely related fields of statistical learning and data mining. To better understand the goals and challenges of machine learning, it is worthwhile to review topics in both statistical learning and data mining.

In general, statistical learning encompasses techniques to reach a better understanding of data [30]. The entire dataset *X* consists of multiple data points $x = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ for p independent attributes. For cases where each x_i data point has a corresponding dependent variable y_i, the techniques from statistical learning can be used to predict or describe this relationship using a technique called supervised training. The goal is to find a function $\widehat{f}$ that approximates a real function f that relates all data points *X* with each output *Y* such that $\widehat{y} = \widehat{f}(\mathrm{X})$. For these problems, the goal is either prediction or inference. Prediction allows for statistical determination of the output given an input that has not been seen before. When a prediction is desired, statistical methods resulting in complex or black-box models are appropriate since it is unnecessary to fully understand the intricacies of the relationship between data points and their output. For inference cases, the details of the relationship function ($\widehat{f}$) are paramount.

Statistical methods also apply to datasets without corresponding output values. In these cases, the goal is to find or describe any relationships that exist between data points.

Han *et al.* [31] similarly describe topics and methods in data mining, of which many are shared with both statistical learning and machine learning. The differences between these fields are mostly in their objectives, while they resemble each other closely in methodology. Data mining is a process involving the discovery of relevant patterns and knowledge from an existing mass of data, while the main objective of statistical learning is to understand relationships in data. Machine learning has a focused objective in generating specific predictions on new data.

Key data-related concerns of machine learning are shared in data mining: data cleaning, integration, selection, and transformation. Data mining seeks to find patterns, associations, and correlations within data. Much like machine learning, data mining uses methods in classification, regression, and clustering to achieve this. There are numerous methods used in data mining to achieve its goals, and these methods are also used frequently in machine learning. Classification methods include decision trees, Bayes classification, rule-based classification, Bayesian belief networks, backpropagation, support vector machines, association rule mining, and lazy learners. Regression methods in data mining are borrowed from statistics and include linear, logistic, and lasso regression. Clustering techniques used in data mining include *k*-means clustering, probabilistic hierarchical clustering, and subspace clustering.

F. Robotics

Robotics is the enabling technology for artificially intelligent systems to interact with their environments [2]. With robotics helping realize the objectives of AI, AI enhances the capabilities of robotics [31]. Improvements in machine vision, prediction models, and deep neural networks allow robots to perceive, plan and execute, improving the autonomy of machines by providing enabling technologies and improving efficiencies with

self-driving cars, interplanetary rovers, drones, and nanorobots as examples

IV. Privacy Concerns With AI

AI technology increases the ability to gather, analyze, and synthesize incredibly vast quantities of data from a variety of sources at a scale never before possible. Further, this type of technology is now widely available to all kinds of individuals and organizations around the globe. While computational technology previously impacted privacy, this new ability greatly increases the potential for privacy harms that were not previously contemplated.

One common method for evaluating privacy harm is to conduct a privacy impact assessment (PIA). Conducting effective PIAs is itself an evolving area of research where current methodologies and frameworks can be borrowed and applied to AI. Guidelines, as in [33], suggest starting a PIA as early as possible and include detailed methodologies. Guidelines from CNIL [34] suggest a compliance-based approach for PIA, where fundamental principles of data subjects' rights are examined to identify and mitigate risks. Steps in a PIA include preliminary analysis, project analysis, privacy analysis, and report generation. Our focus here is on the privacy analysis step and how it could be applied to AI as a whole.

Privacy analysis includes tabulating all functional and technical characteristics of the system. Characteristics can be identified through standardized questionnaires and examples of characteristics include [33]:

1) general characteristics
   a) involves data systems;
   b) involves data sharing;
   c) involves identities of individuals;
   d) involves a change in existing protections;
2) technical characteristics
   a) involves observing or tracking individuals (e.g., video cameras, cell phones, or GPS);
   b) involves communication over untrusted media (e.g., the Internet).

Wright *et al.* [35] compared PIAs across several countries to identify what core features make a PIA effective. They suggested:

1) identify information flows;
2) categorize information, including all types of privacy (Personal, Location, Behavioral, Communication);
3) document how information is processed and by whom;
4) determine how information is stored and secured;
5) list who has access to the information.

Privacy risks deal primarily with collecting and managing data, under the assumption that a system has been implemented with a compliance-based approach. Compliance requirements, such as those regulated by the GDPR, will include stipulations that data subjects must be informed, be given consent, have a right of access, have rights for rectification and erasure, and have a right to object [34].

To assess privacy risk in an AI system, we consider the previous guidelines and apply them to each component of an AI system, where we attempt to qualify whether privacy risk exists or not. The risk here is the likelihood that a threat will impact privacy negatively. Negative consequences are a loss of confidentiality, integrity, or availability of the concerned data.

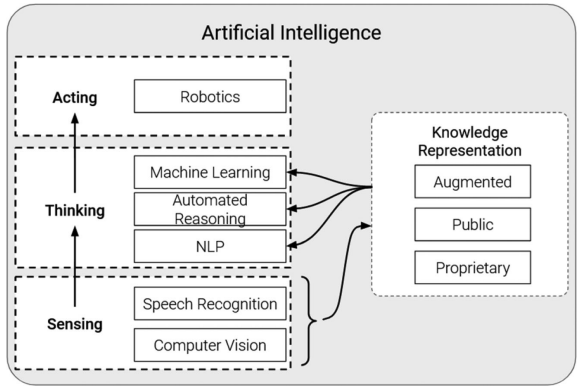

Fig. 2. Functions of AI and information flow.

For a simplified approach using the privacy considerations provided by Wright *et al.* [35], the following two overarching criteria stand out.

1) Does the technology involve information flow? Information flows are concerned with information collection, storage, and processing.
2) What category of information is involved? Consideration of types of data could affect which approaches can be applied to protect privacy, as different categories of data may have unique representations.

We will apply this approach to the aforementioned components of AI systems (automated reasoning, robotics, NLP, computer vision, knowledge representation, and machine learning). As AI continues to evolve and begins to include new technologies, such a process should be repeated as well as a detailed privacy analysis for every implementation of an AI system.

We reorganize the components of the AI system into simplified functional categories to visualize where information flows within an AI system, applying our derived privacy assessment (see Fig. 2). Information is stored in a machine-usable format in some standardized knowledge representation. This includes public information, such as public ontologies for automated reasoning, augmented information, such as labels for training sample sets used in machine learning, and proprietary system-specific data. Sensing through computer vision and speech recognition feeds into the system's information base, which allows the system to think. After processing, or thinking, the system makes a decision and acts (through robotics, if the system has a physical interface).

After considering information flows in AI systems, we find that there are three functional areas that directly involve potentially private information: sensing, knowledge, and thinking. These encompass computer vision, speech recognition, NLP, knowledge representation, automated reasoning, and machine learning and as such, we must consider privacy threats that exist in these areas.

A. Computer Vision and Speech Recognition

There is a large base of research into privacy in computer vision, and one common element is that consideration to privacy is applied to the representation of the visual information in data [36]–[40]. Privacy issues with computer vision range from age estimation, occupation recognition, face recognition, and social relationships [39]. Wu *et al.* outlined the dilemma of computer vision in [36]. To assist humans, artificially intelligent systems need to be able to recognize details of their environment without collecting private information. Private information can be leaked directly or inferred, such as revealing your location by having imagery with landmarks nearby [39].

Speech recognition, much like computer vision, has had major advances corresponding to the application of deep learning methods [41]. Threat models for speech recognition focus on compromising voice data, which can allow adversaries to determine mental health, stress, smoking habits, and other conditions [42].

Solutions to privacy concerns with speech recognition and computer vision involve software-based approaches or algorithms running on the data, as in [38], [40], and [43]. Other concerns of speech recognition, such as those discussed in [44], include voice-only interfaces that would require users to speak private information aloud.

B. Natural Language Processing

Tang *et al.* [45] describe the purpose of natural language processors as deriving intent from human requests, represented in textual form. After the speech is recognized, determining meaning from the sequence of words is considered a classification problem. Privacy issues arise in applications sensitive in nature, such as call automation for healthcare institutions. Records created from speech recognition to be analyzed in NLP will likely contain private information. The difference between NLP databases and typical data stores of private healthcare information is that speech is not necessarily received in a structured format, and addressing privacy issues is not straightforward.

Privacy issues in NLP are addressed using machine learning techniques. These techniques can be used for intelligently redacting private information, as described in [46], or by creating synthetic data, as described in [47].

C. Knowledge Representation

The information available to AI systems is immense and consists of data representations of images, videos, speech, attributes, categorizations, and more. This knowledge base has the potential to contain private information on many individuals. Data processing entities of all kinds tend to develop privacy compliance programs in response to this risk. These programs are generally legally focused on the mitigation of organization risks related to noncompliance with the plethora of privacy regulations around the globe. By design, such programs are not intended nor serve the individual data subject, nor inform them as to the actual privacy they may have in the use of a given technology.

In [48], imbalances between data processing entities and individuals are highlighted. Privacy risks in knowledge representation materialize in automated data analysis and data breaches.

In a literature review by Smith *et al.* [11], categorizations of privacy concerns were surveyed. Top concerns for information privacy are large amounts of data collected, unauthorized secondary use (by the collecting organization or by external organizations), data misrepresentations (lack of quality), poor decision making from available data, and data aggregation.

Large amounts of data collection is a common concern shared among consumers. Attributes regarding individual personalities, characteristics, and actions are collected, violating privacy. *Unauthorized secondary use* (internal or external) is a privacy risk where a user may consent to an organization collecting data, but that data are subsequently sold off to a third party or otherwise used for a purpose other than how it were consented to be used. *Data misrepresentation* is the privacy risk that information regarding an individual could be incorrect or inaccurate. *Poor decisions* based on collected information is a concern that the narrow representation of an individual in data may lead systems to reach incorrect decisions, such as automating credit approvals. Human interaction may allow for more meaningful decision making if the data and/or system is not robust. *Data aggregation* is the risk where disparate data sources may not contain enough information to reveal any private information individually, but their combination is enough to pose a privacy concern.

D. Automated Reasoning

Automated reasoning is the ”common-sense” aspect of AI systems. These are built upon ontologies that provide a basis for AI to reason and make inferences autonomously. These common-sense repositories, as discussed in [27], will typically draw from public sources, such as Wikipedia or Google Search [28], and will have limited privacy impact.

Ambiguities on the boundaries of public information, however, can lead to privacy concerns. Facial recognition software built on publicly available data (e.g., social media profiles), as with Clearview AI, has lead to an outcry against, and legal ramifications for such systems [49]. These information sources, while built on ”public” data, cannot be guaranteed to comply with various international privacy laws and could yield incorrect results. A misclassification could have a harmful impact, for example, if used in a law enforcement scenario, leading to wrongful arrests.

E. Machine Learning

Machine learning lies at the core of most AI operations. From processing visual data, speech, and making decisions, machine learning is critical. Al-Rubaie and Chang [50] outlined common privacy threats in machine learning systems. Five risks are identified, which can be loosely categorized based on the role of machine learning in the realization of the risk. These categories are: 1) risks that are introduced by implementing a machine learning system, targeting the system itself; 2) new risks introduced to existing systems enabled by machine learning; and 3) risks to machine learning systems introduced by using other

machine learning systems. We will explore privacy concerns in these three scenarios.

1) *Machine Learning as the Target:* These privacy risks are exploits of vulnerabilities in the design, implementation, or execution of machine learning systems. Numerous threats exist, and the most commonly cited ones are described here.

Adversarial examples are inputs into a machine learning model designed to return an incorrect classification. Kurakin *et al.* [51] outline different methods to create adversarial examples: first gradient sign, one-step target class, basic iterative, and iterative least-likely class. A prime example of adversarial examples is efforts to bypass email spam filters.

Model reconstruction, or model extraction, is an attack that attempts to reverse engineer the operating parameters of various ML models, including linear regressions, neural networks, and decision trees. Tramer *et al.* [52] showed that such attacks could be used to bypass fees associated with using machine learning as-a-service platforms, evade ML filtering models (e.g., e-mail spam filters), or violate privacy-sensitive training data. Tramer *et al.* showed that exploiting confidence values along with classification labels allows for the deduction of various ML models with a relatively small number of queries, allowing for the replication of ML models in a matter of minutes.

Attribute reconstruction attacks attempt to reconstruct the input attributes that are represented as a feature vector for machine learning models that make these feature vectors visible [50] (i.e., allowing for an attacker to determine attributes based on the feature vector).

An example of an attribute reconstruction attack is shown by Gambs *et al.* [53], specifically targeting decision trees. In this demonstration, the original contributing dataset for a decision tree is partially recreated. The consequences of this attack could reveal private information, such as the reconstruction of a fingerprint image from a fingerprint feature vector.

Model inversion attacks are similar to attribute reconstruction attacks, except for situations where information about the feature vector is not known (i.e., allowing for an attacker to determine the feature vector based on the output of the model). In these situations, instead of manipulating the feature vector directly, an adversary can manipulate input directly and evaluate the classification confidence for each input. In a hill-climbing attack, subsequent inputs are modified with the aim of increasing the confidence value returned by the model.

Fredrikson *et al.* [54] demonstrated potential for a model inversion attack, targeting decision trees and neural networks. This attack could reveal private information depending on the purpose of the machine learning model. For example, with a computer vision system designed to provide identification through facial recognition, a model inversion attack would be able to partially recreate an image of an individual's face given their name. Their testing resulting in 75% accuracy with an 87% identification rate.

2) *Machine Learning as the Tool:* These are privacy risks introduced through the use of machine learning. In these, the extensive predictive capabilities of machine learning systems are abused by adversaries to learn private information that would otherwise be difficult to deduce.

Attribute *inference attacks are designed to learn the private attributes* of an individual. This form of attack can reveal information such as location, gender, ethnicity, and political affiliation [55], [56]. Information to use in attribute inference can come from various publicly available sources, such as social media, recommender systems, and mobile platforms [55]. These attacks use machine learning systems trained on feature vectors of publicly available information with a set of known private information from training samples to infer private information of individuals outside the training set. Jia and Gong [55] showed that not only can an individual's publicly available information be used to infer private information about themselves, so too can the public information of their social connections be used against them. Jia *et al.* showed that using public Twitter information ("tweets") and analyzing a user's friends' tweets is equally viable to yield the same private information with nearly identical accuracy. This technique could be used to reveal directly sensitive information or to reveal other attributes that could be combined to perform a reidentification attack.

Re-identification or deanonymization is the process of discovering the identity of an individual who contributed data that subsequently had anonymization techniques applied, such as generalization, masking, and k-anonymity [48]. As described by Narayanan and Shmatikov [57], [58], datasets that are large and sparse are especially vulnerable to reidentification attacks. Such datasets are characterized by having many dimensions (large) with relatively few dimensions being populated for any given record (sparse).

With the increasing availability of data, the threat of reidentification is similarly increasing. Even with anonymization and perturbation techniques applied, high-dimensional data can still be reidentified, noting that only 33 bits of information can precisely identify an individual [58]. Reidentification attacks have been demonstrated across multiple domains, including social networks, genetic data, Internet activity, and writing style.

The task for reidentification on large sparse datasets can be framed as a problem for machine learning, as shown in [59] and [60]. In [59], Kosinski *et al.* trained a machine learning model on 58 000 volunteer questionnaires along with their Facebook profiles, including all of their "likes" and activity. Attributes collected from volunteers included private and sensitive information, including sexual orientation, ethnicity, religious and political views, age, and gender. By training their model to find a relationship between like patterns (a large, sparse dataset) and personal or private information, the resulting model showed high accuracy for binary classifications (e.g., drug usage, democrat or republican), with up to 93% accuracy for predicting a user's gender. Numeric quantities were less accurate between having 17% accuracy for predicting a user's satisfaction with life and 75% accuracy for predicting a user's age.

Another demonstration by Sharad and Danezis [60] used anonymized phone call graphs released in Orange's "Data for Development" challenge. In this attack, a larger call graph was anonymized into small subgraphs (or Egonets) with only some original features of call information retained. By training a machine learning model on the graph features of node degree and neighbor node distribution, Sharad and Danezis showed that the

nodes within the subgraphs can be reidentified with up to 90% accuracy.

An important contributing factor to reidentification attacks is that all data must have some utility. This implies that there must be some unique features between records, else there is little use. These small differences can further be augmented with auxiliary information (such as survey data as in the Facebook study or separate subgraphs as in the Orange study) to successfully discover identities or sensitive attributes.

3) *Machine Learning as the Target and Tool:* These attacks leverage the power of machine learning systems to target weaknesses in other systems.

A membership *inference* [61] *attack attempts to determine if a known* sample input belongs to the training set of a machine learning model. This attack could reveal private information about individuals with models that are sensitive in nature, such as those that provide health-related classifications, as in [62].

For example, if an employer is deciding to hire an applicant or not, a membership inference attack would allow the employer to use the applicant's information to determine how likely they are to have contributed to a separate study on cancer patients. The employer may then use this information to influence their hiring decision.

Membership inference attacks are enacted by an adversary that creates separate machine learning models mimicking the behavior of the target model. These adversarial models, called shadow models, are trained on similar data to the target model, called shadow data, which can be derived from various sources, such as general statistics of the population, semisupervised learning with actual samples, queries on the target model, or region-based targeting [63]. By training several shadow models on differing shadow data, an attack model is trained to return "in" or "out" given a sample that was either used or not used in training each shadow model.

This attack works due to the nature of the output of machine learning systems, often being a prediction vector of potential classifications along with the probabilities for each classification. The classification with the highest probability is then taken to be the result of the model. As detailed by Yeom *et al.* [64], a major factor in conducting a membership inference attack is overfitting, which is a flaw in how machine learning models are trained. Yeom *et al.* further describe mechanisms to construct a membership inference attack without the use of a secondary attack model.

A machine learning model is overfitting when the generalization error for data points the model was trained on is significantly less than samples from the general distribution of all data points. In other words, if a model is much more confident in its predictions for samples it has seen than for samples it has not seen, then that model is said to be overfit.

Shokri *et al.* [61] show an example of this attack, specifically targeting machine learning-as-a-service providers. In these scenarios, an adversary can create replicas of the target model and train them on potentially completely separate data drawn from the same statistical population of the target model. The success of these attacks ranged from 74% to 94%.

TABLE I
AI PRIVACY RISKS WITH SOME MITIGATIONS

AI Component	Mitigation Strategies
Knowledge Representation	Cryptography, Perturbation, Anonymization
Natural Language Processing	Redaction, Synthesis [47]
Automated Reasoning	Human-in-the-Loop
Machine Learning	Cryptography, Perturbation, Dimensionality reduction

V. MITIGATIONS OF PRIVACY RISKS

The Privacy-AI boundary is best understood as a moving target, as rules and regulations for privacy adapt to societal norms and expectations while the components of an artificially intelligent system evolve. Mitigating privacy risks stemming from AI can be accomplished by addressing the risks associated with each component of AI systems in the context of current privacy expectations. The components of concern will vary as the state of the art evolves. As previously discussed, high-risk areas for privacy in the current state of AI are knowledge representation, NLP, automated reasoning, and machine learning. Table I summarizes general mitigation strategies, of which a selection will be explored in more detail within this section.

A. General Principles

Protecting private information in any system is a common issue that has been researched extensively. The main areas of concern are data storage, transfer, and processing [65]. Implementing systems that are sensitive to privacy expectations and requirements is discussed in detail by Langheinrich [66]. Generally, the recommended approach involves selecting appropriate privacy-enhancing technologies (PETs) that meet requirements under some predefined threat model. Instances that aim to provide, for example, improved customer experiences may adopt a threat model where suboptimal PETs are adopted, while another system, such as a healthcare web portal, could require more stringent controls. Spiekerman and Cranor provide a framework that describes system characteristics for differing degrees of user identifiability across four privacy stages. Lower levels of privacy are characterized by reliance on notice and consent or privacy by policy. The highest levels of privacy guarantees are only possible through integrating system design with privacy goals [65].

An overview of this notion of privacy by design is offered by Langheinrich [66], who suggested that there is an overlap between advances in information technology security and privacy. Langheinrich posits that adequate security measures will, as a by-product, provide privacy. However, security-based approaches should not be relied upon exclusively, and the security principles discussed previously should be adhered to, such as those provided in [67] and [68]. As will be shown in the following sections, many PETs stem from security research.

B. Protecting Knowledge Representations

The objective of privacy preservation is to limit the risk of disclosing sensitive data, where sensitive data are referring to

either identity information or a confidential attribute. Privacy preservation techniques should avoid publishing original data and also maintain data utility.

There are varied and related fields of research for privacy preservation in information systems that follow the goal of protecting the underlying data. From [13], these fields include privacy-preserving data mining and statistical disclosure control, and anonymization.

SDC and PPDP have similar objectives, with SDC focusing on utility over privacy and PPDP focusing on privacy over utility [13]. Descriptions of solutions, techniques, models, and algorithms pertaining to these areas of research follow.

There are two broad approaches for achieving privacy in data: cryptography and perturbation [50], [69]. Techniques in cryptography can be used to reversibly transform private information, preventing an adversary from learning any details of the data. Cryptographic techniques are effective up to the strength of the implemented protocol. Perturbation techniques make irreversible changes to values in data and are be applied following a privacy-enhancing protocol that balances a tradeoff between data utility and privacy. Techniques in perturbation aim to reduce the accuracy of private information to the point it no longer reveals information true to an individual while still yielding some desired statistical accuracy on the data as a whole. The effectiveness of perturbative methods can be captured by measuring information loss against the disclosure risk, as described in [13].

Cryptographic techniques that can be applied to knowledge representations are described in [50], [70], [71], and [72]. The most prevalent protocol mentioned for managing computations on data while maintaining privacy is oblivious transfer, which is used as a building block for implementing other protocols. More complex protocols include homomorphic encryption, Yao's garbled circuits [73], and secure processors.

Homomorphic Encryption is a branch of encryption that allows for meaningful computations on encrypted data. In traditional encryption applications, computations such as addition, multiplication, subtraction, and comparisons can only be applied to plaintext values. This means that before computation, encrypted and potentially sensitive or private data must be decrypted before being computed upon. Decrypting data create vulnerabilities within information systems, as the data can then potentially be stolen. Homomorphic cryptosystems solve this vulnerability by removing the necessity for data to be decrypted before being operated on.

Several encryption schemes have been proposed that enable homomorphic encryption. These schemes can broadly be divided into those that are partially homomorphic and those that are fully homomorphic. With a definition from [74], a given encryption scheme is homomorphic over an operator if, given two operands, the encrypted result of the operation on the unencrypted operands is equal to the operation on the encrypted operands or as a simplified expression (using multiplication as an example)

$$E\left(o_1 \cdot o_2\right) = E\left(o_1\right) \cdot E(o_2).$$

Partially homomorphic schemes allow for either addition or multiplication, while fully homomorphic schemes allow both. Popular partially homomorphic schemes include RSA, El-Gamel, and Pallier, while schemes from Brakerski/Fan-Vercauteren and Cheon-Kim-Kim-Song are among the more popular fully homomorphic schemes [74], [75].

Fully homomorphic encryption schemes have been implemented in programming libraries and applied to privacy-related issues, as discussed in the work of Vizitiu *et al.* [76]. From [76], homomorphic encryption was applied to privacy-sensitive medical data. Data were encrypted using the fully homomorphic scheme MORE [77]. The results of Vizitiu *et al.* showed that training machine learning models on encrypted data representations are possible, with emphasis that their application required selection of a weaker encryption scheme.

Garbled circuits are a mechanism that allows two parties to perform a computation on their inputs without revealing each party's input to the other, as described in [78]. When using garbled circuits, one party acts as the garbler, who is primarily responsible for encryption, and the other as the evaluator, who runs the encrypted inputs over the encrypted circuit. The garbler creates an encrypted representation of a circuit and sends this encrypted circuit to the evaluated along with their own encrypted input. Using cryptographic protocols (e.g., oblivious transfer), the second party is able to receive the encrypted value for their own input from the garbler without revealing their unencrypted input.

A use case for garbled circuits is given by Sadeghi *et al.* [79], in providing privacy for a facial recognition application. The use of the system does not want to reveal the face which they are sending as a query, and the system should not reveal anything about the faces in its database. The authors find an improved method of implementing garbled circuits in a hybrid approach alongside homomorphic encryption.

Secure Processors create an additional level of security in information systems right at the processor level. Requiring dedicated hardware, such as Intel SGX processors [50], secure processors reduce the vulnerability of data leakage by creating additional requirements in processing instructions, such as instruction and bus encryption. An overview and comparison of secure processors are given by Sau *et al.* [80], who compared processor implementations (hardware versus software), performance, execution isolation, secure storage, remote attestation, secure provisioning, and trusted path characteristics of various secure processors.

An application of secure processors is given in by Dang and Chang [81], where privacy concerns regarding cloud storage of data are explored. In their experiment, it is assumed that a cloud provider desires to implement data deduplication while preserving the privacy of data that is submitted to be stored. In their proposed architecture, Dang and Change utilize secure processors to "seal" data securely, protecting any privacy-sensitive information that may be submitted.

Perturbative methods described in [13] include noise addition, data/rank swapping, and microaggregation. These methods, combined with some nonperturbative techniques (sampling,

generalization, coding, and suppression) are used in more complex privacy models that provide anonymity in data.

The state of anonymity implies that no data can be linked back to an individual through directly identifying attributes or combinations of attributes that function as quasi-identifiers. Prevalent anonymizing models are *k*-Anonymity, *l*-Diversity, *t*-Closeness, and differential privacy (DP).

k-Anonymity and related models are regarded as sufficient techniques to protect privacy, implemented through approximation due to the complexity of an optimal solution [13], [71], [82]. The k-Anonymity model creates a guarantee of privacy within a set of data such that no individual record can be determined with a greater probability than 1/*k*. This is accomplished by generalizing all quasi-identifying attributes to be identical within groups of size *k*. A weakness of k-Anonymity is if the sensitive attributes are not varied, the quasi-identifier generalization will fail in providing any privacy. For this case, *l-Diversity* is an improvement upon k-Anonymity that maintains a distribution of values within each grouping. A further improvement on *l*-Diversity is *t-Closeness*, which is satisfied if the distribution of sensitive attributes within each k-group is within threshold *t* of the global distribution of those attributes.

In a study focusing on how data protected with k-Anonymity would impact a larger system, Wimmer and Powell [83] explored the impact of *k*-Anonymity on machine learning. After applying *k*-Anonymity to several datasets, Wimmer and Powell ran various machine learning algorithms, including neural networks, decision trees, and Bayesian classifiers. Their results show that some algorithms maintain similar performance levels on anonymized data.

Perturbation techniques from Al-Rubaie and Chang [50] include DP, local DP (LDP), and dimension reduction (DR). Aggarwal and Xu [71] listed randomization as a viable technique, but also note inadequacy in protecting outliers in data.

DP is a mechanism that introduces noise into a computation system, either on the input, output, or at an intermediate step. In DP, a probabilistic threshold of privacy is decided that will dictate how closely the actual resulting output of a query on data resembles the real data without noise added. The actual output of a differentially private system is indistinguishable from the raw output with any one input withheld.

Formally, the definition of DP is given in [84]. A randomized function, K, provides privacy up to a parameter ε for two datasets *D* and *D*' (with Hamming distance $d(D,D') \leq 1$, and $S \subseteq$ Range*(K)* if

$$\Pr\left[K\left(D\right) \in S\right] \leq e^{\varepsilon} \times \Pr\left[K\left(D\prime\right) \in S\right].$$

Zhu *et al.* explored the effectiveness of DP in various AI in [85]. They found that DP can be applied to protect the privacy of individuals in data by hiding an individual's contribution to a dataset in the aggregation of that dataset. Further, Zhu *et al.* found that differentially private datasets can be used when applying machine learning or deep learning algorithms, with promising results.

LDP is the application of DP to local inputs into a machine learning system, with the goal of introducing plausible deniability of each contributing party to the model, with one popular implementation being randomized response.

Dimensionality Reduction (DR) is a technique to reduce the number of dimensions in a dataset by projecting upon a lower dimension hyperplane. This transformation is lossy and irreversible. There are a number of techniques that can be applied to achieve: 1) random matrix, 2) principal component analysis, 3) discriminant component analysis, and 4) multidimensional scaling. Even though the techniques are irreversible, there is still the possibility of approximating original inputs. As an added protection, DR can be combined with DP.

Applying PCA is described by Shlens in [86]. The goal of PCA is to determine if there is an alternative basis for collected *m*-dimensional data that provide a better representation than the basis on which the data were collected. In other words, Shlens offered that the goal is to find P such that $XP = Y$, with an original $m \times n$ matrix *X*, rebased matrix *Y*, where *Y* provides a more statistically meaningful representation of *X*.

PCA has been used in AI applications, as with Yang *et al.* [87], who applied the method in a system to preserve privacy in facial recognition applications. Their results showed that privacy can be preserved in such applications and the privacy of participants can be preserved.

C. Protecting NLP

As mentioned in [47], there are two approaches to preserving privacy in NLP: redaction and synthesis. The unique issue in NLP differing from privacy risks with stored data (as in knowledge representations) is that speech data are typically unstructured. Privacy models designed for structured data cannot be easily applied.

Approaches to redaction are described in [45] and [46]. One simple approach to redaction is through *named entity extraction*. After finding named entities, *distortion*, *dissociation*, or *value-class replacement* can be used to redact sensitive information. Distortion obfuscates real values with meaningless data. Dissociation shuffles real values of entities around. Value-class replacements change entity values with their category (e.g., exchanging "Bob" for "PERSON").

Synthesis involves generating artificial text with statistical significance to real samples such as drawing samples from a Bayesian posterior predictive distribution [47].

Medical records are often subject to scrutiny for privacy-sensitive operations, such as data processing. In [88], Sadat *et al.* showed that hospital records can be privately shared following a framework that includes an NLP privacy-preserving step of identifying and removing sentences containing low-probability bigrams. In a scheme that includes homomorphic encryption, Sadat *et al.* showed that the privacy of medical datasets can be preserved using NLP-based techniques while maintaining analytical significance.

D. Protecting Automated Reasoning

Privacy concerns with automated reasoning relate to the quality of information used for decision support. In the scenario of misclassification based on inaccurate information, mitigation comes in the form of human-in-the-loop (HITL). HITL is a technique more closely related to machine learning and suggests that classification is improved (or, only possible) when a human

is involved [89]. Requiring a human to review the classification decided based on available information in a decision step, such as automated reasoning, helps to ensure that the system is not behaving in a way counter to the process it is intended to be supporting.

In the context of AI, Rahwan [90] expanded on the HITL concept to include a societal aspect with consideration for the fact that the outcomes of AI systems are far-reaching. The proposed system, society-in-the-loop considers inputs from society-at-large as a factor in the role of a human supervisor in an AI system.

In analyzing an AI system and attempting to determine where and to what extent a human should be involved, Shneiderman [91] offered an approach to assess various types of AI systems. For concerns from too little or too excessive human interaction, Shneiderman's human-centered AI framework offered a process that supports various goals, including security, resilience, and privacy.

E. Protecting Machine Learning

The attack surface of machine learning is outlined in [92] and includes input features, processing, results and output, and communicating results. Machine learning poses further privacy risks if it is used as a tool to invade the privacy of individuals. We will address concerns of privacy risks attributed to machine learning itself and risks that machine learning creates if used inappropriately.

Protecting machine learning: Privacy risks are introduced when adopting machine learning systems that involve private information. A number of attacks exist that exploit weaknesses in the attack surface of machine learning systems, such as those previously described. Mitigations for specific attacks are briefly discussed here. To address adversarial examples, *adversarial training* is a technique to use purposeful adversarial examples in the training phase of machine learning to make models more robust against malicious adversarial examples.

In adversarial training, a generative adversarial network is used to create generator (G) model and discriminator (D) model. G works to create samples that fool D, while D learns to differentiate samples from G and real samples. Han *et al.* described how this approach can be applied to the application in affective computing and sentiment analysis.

Other techniques for protecting machine learning from being stolen completely is a topic discussed in [52]. Effective countermeasures are: *reducing or removing confidence values* returned by the model or applying *DP* to the training data.

Guarding against model inversion is outlined in [54]. Simply, *reducing the quality* of information machine learning returns (in the case of neural networks) or the *order in which information is returned* (in the case of decision trees) lowers the effectiveness of these attacks. *DR* is also an approach to reducing the risk of model inversion [50].

Protection against membership inference attacks is discussed in [41] and can be addressed through several efforts. Reducing overfitting, for which *regularization* techniques can be used, *restricting prediction vectors,* or *increasing entropy in prediction vectors* will all help protecting against membership inference. Additional mitigation techniques are discussed in [63] and techniques include model hardening and *API Hardening*, following best practices for API security.

1) *Protecting against machine learning*: Machine learning can be used maliciously to infringe on the privacy of individuals. Attacks include attribute inference and reidentification. Methods addressing these risks are discussed here.

Inferring information or outright reidentifying individuals from data relies on relationships between attributes in the underlying data. Protecting the underlying data to prevent machine learning from revealing private information is discussed in [57] and [93] with the techniques of *k-Anonymity* and *DP* being examples. Care should be taken in applying anonymizing techniques, as these techniques make a tradeoff between privacy and utility. Attribute reconstruction attacks, as described in [53], have an effect that relies on the released classifier. Efforts similar to those in protecting against membership inference will similarly apply here, such as *reducing the output accuracy*.

VI. Conclusion

Privacy concerns surrounding AI are a critical topic that should be taken seriously. As the capabilities of computer systems grow, so do the capabilities of AI systems. Affronts to privacy in any form are also an affront to human rights and democratic values [94].

The challenge of assessing privacy risks within AI is the adaptive and evolving nature of the field, with AI being the result of aggregating disparate fields in an effort to create something greater than the parts. Our research suggests that there is no singular solution to such a problem and proposes a general approach to evaluating the components of AI against localized privacy expectations using established PIA methodologies.

To preserve privacy in AI systems, each component of the system must be considered against privacy frameworks individually. In the current state of AI, privacy risks are dominant in knowledge representations, NLP, automated reasoning, and the creation, operation, and use of machine learning models. Removing privacy risks from AI systems is not a simple task, and is an effort that will continuously evolve as AI similarly progresses. Thus, future work in the field of privacy in AI would benefit from continuous reassessment of how constituent technologies contribute to privacy risks, and how they interact with each other technology in an AI system.

References

[1] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. London, U.K.: Oxford Univ. Press, 2014.
[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Kuala Lumpur, Malaysia: Pearson Education, 2016.
[3] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
[4] D. Walsh, J. M. Parisi, and K. Passerini, "Privacy as a right or as a commodity in the online world: The limits of regulatory reform and selfregulation," *Electron. Commerce Res.*, vol. 17, no. 2, pp. 185–203, 2017.

[5] G. Fowler, "Perspective—Alexa has been eavesdropping on you this whole time," May 2019. [Online]. Available: https: //www.washingtonpost.com/technology/2019/05/06/alexa-has-been-ea vesdropping-you-this-whole-time/

[6] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, ." *J. Pers. Social Psychol.*, vol. 114, no. 2, 2018, Art. no. 246.

[7] Oxford Living Dictionaries, "Privacy." [Online]. Available: https://en.oxforddictionaries.com/definition/us/privacy, Accessed: Sep. 21, 2018

[8] S. T. Margulis, "On the status and contribution of westin's and altman's theories of privacy," *J. Social Issues*, vol. 59, no. 2, pp. 411–429, 2003.

[9] D. J. Solove, "A taxonomy of privacy," *Univ. Pennsylvania Law Rev.*, vol. 154, no. 3, pp. 477–564, Jan. 2006.

[10] M. Kwasny, K. Caine, W. A. Rogers, and A. D. Fisk, "Privacy and technology: Folk definitions and perspectives," in *Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2008, pp. 3291–3296.

[11] 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, 1996.

[12] F. Belanger and R. E. Crossler, "Privacy in the digital age: A review of information privacy research in information systems," *MIS Quart.*, vol. 35, no. 4, pp. 1017–1042, 2011.

[13] J. Domingo-Ferrer, D. Sanchez, and J. Soria-Comas, "Database anonymization: Privacy models, data utility, and microaggregation-based inter-model connections," *Synth. Lectures Inf. Secur., Privacy, Trust*, vol. 8, no. 1, pp. 1–136, 2016.

[14] T. Dalenius, "Towards a methodology for statistical disclosure control," *Statistisk Tidskrift*, vol. 15, pp. 429–444, 1977.

[15] P. A. Norberg, D. R. Horne, and D. A. Horne, "The privacy paradox: Personal information disclosure intentions versus behaviors," *J. Consum. Affairs*, vol. 41, no. 1, pp. 100–126, 2007.

[16] S. Kokolakis, "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon," *Comput. Secur.*, vol. 64, pp. 122–134, 2017.

[17] D. Poza, "11 of the worst data breaches in media," Feb. 2019. [Online]. Available: https://auth0.com/blog/11-of-the-worst-data-breaches-in-me dia/

[18] A. Acquisti, I. Adjerid, and L. Brandimarte, "Gone in 15 seconds: The limits of privacy transparency and control," *IEEE Secur. Privacy*, vol. 11, no. 4, pp. 72–74, Jul./Aug. 2013.

[19] P. Stone *et al.*, "Artificial intelligence and life in 2030," One Hundred Year Study Artif. Intell., Rep. 2015-2016 Study Panel, Stanford Univ., Stanford, CA, USA, 2016, Art. no. 52.

[20] D. Vernon, *Artificial Cognitive Systems: A Primer*. Cambridge, MA, USA: MIT Press, 2014.

[21] F. Van Harmelen, V. Lifschitz, and B. Porter, *Handbook of Knowledge Representation*, vol. 1. Amsterdam, The Netherlands: Elsevier, 2008.

[22] Britannica, "Laws of thought," Apr. 2019. [Online]. Available: https://www.britannica.com/topic/laws-of-thought

[23] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, 2018, Art. no. 7068349.

[24] P. Tanwar, T. Prasad, and M. S. Aswal, "Comparative study of three declarative knowledge representation techniques," *Int. J. Comput. Sci. Eng.*, vol. 2, no. 7, pp. 2274–2281, 2010.

[25] J. Ramadas and V. N. P. Marg, "Lecture notes: Knowledge representation techniques," Sep. 2002. [Online]. Available: http: //www.hbcse.tifr.res.in/jrmcont/notespart1/node38.html

[26] A. Bundy, "A survey of automated deduction," in *Artificial Intelligence Today*. Berlin, Germany: Springer, 1999, pp. 153–174.

[27] "Teaching machines common sense reasoning," Oct. 2018. [Online]. Available: https://www.darpa.mil/news-events/2018-10-11

[28] T. H. Trinh and Q. V. Le, "A simple method for commonsense reasoning," *arXiv preprint arXiv:1806.02847*, 2018.

[29] H. J. Levesque, E. Davis, and L. Morgenstern, "The winograd schema challenge," in *Proc. AAAI Spring Symp. Ser.*, pp. 552–561, 2011.

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction To Statistical Learning*, vol. 112. Berlin, Germany: Springer, 2013.

[31] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.

[32] J. A. Perez, F. Deligianni, D. Ravi, and G.-Z. Yang, "Artificial intelligence and robotics," *arXiv preprint arXiv:1803.10813*, 2018.

[33] Information and Privacy Commissioner of Ontario, "Planning for success: Privacy impact assessment guide," May 2015.

[34] CNIL, "Privacy impact assessment (PIA) methodology," Feb. 2018.

[35] D. Wright, R. Finn, and R. Rodrigues, "A comparative analysis of privacy impact assessment in six countries," *J. Contemporary Eur. Res.*, vol. 9, no. 1, pp. 160–180, 2013.

[36] Z. Wu, Z. Wang, Z. Wang, and H. Jin, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 606–624.

[37] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition: Privacy implications in social media," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 19–35.

[38] P. Speciale, J. L. Schonberger, S. B. Kang, S. N. Sinha, and M. Pollefeys, "Privacy preserving image-based localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5493–5503.

[39] T. Orekondy, B. Schiele, and M. Fritz, "Towards a visual privacy advisor: Understanding and predicting privacy risks in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3686–3695.

[40] S. Avidan and M. Butman, "Blind vision," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 1–13.

[41] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1310–1321. [Online]. Available: http: //doi.acm.org/10.1145/2810103.2813687

[42] R. Aloufi, H. Haddadi, and D. Boyle, "Emotionless: Privacy-preserving speech analysis for voice assistants," , *arXiv preprint arXiv:1908.03632*, 2019.

[43] M. S. Barhm, N. Qwasmi, F. Z. Qureshi, and K. El-Khatib, "Negotiating privacy preferences in video surveillance systems," in *Proc. Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst.*, 2011, pp. 511–521.

[44] V. G. Motti and K. Caine, "Users' privacy concerns about wearables," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, 2015, pp. 231–244.

[45] M. Tang, D. Hakkani-Tur, and G. Tur, "Preserving privacy in spoken language databases," in *Proc. Int. Workshop Privacy Secur. Issues Data Mining, ECML/PKDD*. Citeseer, pp. 27–36, 2004.

[46] Y. Li, T. Baldwin, and T. Cohn, "Towards robust and privacy-preserving text representations," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 25–30.

[47] A. G. Ororbia II, F. Linder, and J. Snoke, "Privacy protection for natural language: Neural generative models for synthetic text data," *arXiv preprint arXiv:1606.01151*, 2016.

[48] G. Danezis *et al.*, "Privacy and data protection by design-from policy to engineering," *arXiv preprint arXiv:1501.03726*, 2015.

[49] R. Berman, "This company scraped social media to feed its ai facial recognition tool. Is that legal?" Feb. 2020. [Online]. Available: https://bigthink.com/technology-innovation/facial-recognition-clearvie w?rebelltitem = 4#rebelltitem4

[50] 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.

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

[52] F. Tramer, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing' machine learning models via prediction apis," in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 601–618.

[53] 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.

[54] 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.*, 2015, pp. 1322–1333.

[55] J. Jia and N. Z. Gong, "Attriguard: A practical defense against attribute inference attacks via adversarial machine learning," in *Proc. 27th USENIX Secur. Symp.*, 2018, pp. 513–529.

[56] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors," in *6th Int. AAAI Conf. Weblogs Social Media*, pp. 387–390, 2012.

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

[58] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse 1282 datasets: A decade later," 2019. [Online] Available: https://www.semanticscholar.org/paper/Robust-de-anonymization-of-large-sparse-datasets-%3A-Narayanan-Shmatikov/f41ef0fe589fdfbfe22c1ac5629638773f8d9fe9.

[59] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad.*

Sci. USA, vol. 110, no. 15, pp. 5802–5805, 2013. [Online]. Available: http://www.jstor.org/stable/42590309

[60] K. Sharad and G. Danezis, "An automated social graph deanonymization technique," in *Proc. 13th Workshop Privacy Electron. Soc.*, 2014, pp. 47–58.

[61] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 3–18.

[62] A. Wade, "How artificial intelligence is revolutionising medical diagnostics," Jun. 2019. [Online]. Available: https://www.theengineer.co.uk/ai-medical-diagnostics/

[63] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Demystifying membership inference attacks in machine learning as a service," *IEEE Trans. Serv. Comput.*, to be published, doi: 10.1109/TSC.2019.2897554.

[64] 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.*, 2018, pp. 268–282.

[65] S. Spiekermann and L. F. Cranor, "Engineering privacy," *IEEE Trans. Softw. Eng.*, vol. 35, no. 1, pp. 67–82, Jan./Feb. 2008.

[66] M. Langheinrich, "Privacy by design—Principles of privacy-aware ubiquitous systems," in *Proc. Int. Conf. Ubiquitous Comput.*, 2001, pp. 273–291.

[67] General Assembly Resolution 68/167, "The right to privacy in the digital age," United Nations, Jan. 2014. [Online]. Available: http://undocs.org/A/RES/68/167

[68] OECD, "The OECD privacy framework," 2013. [Online]. Available: http://www.oecd.org/sti/ieconomy/privacy-guidelines.htm, Accessed Nov. 25, 2019

[69] K. Xu, H. Yue, L. Guo, Y. Guo, and Y. Fang, "Privacy-preserving machine learning algorithms for big data systems," in *Proc. IEEE 35th Int. Conf. Distrib. Comput. Syst.*, 2015, pp. 318–327.

[70] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining," *ACM SIGKDD Explorations Newslett.*, vol. 4, no. 2, pp. 12–19, 2002.

[71] C. C. Aggarwal and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms," in *Privacy-Preserving Data Mining*. Berlin, Germany: Springer, 2008, pp. 11–52.

[72] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.

[73] A. C. Yao, "How to generate and exchange secrets," in *Proc. 27th Annu. Symp. Found. Comput. Sci.*, 1986, pp. 162–167.

[74] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes: Theory and implementation," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–35, 2018.

[75] S. Ramesh and M. Govindarasu, "An efficient framework for privacy preserving computations on encrypted IoT data," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8700–8708, Sep. 2020.

[76] A. Vizitiu, C. I. Nita, A. Puiu, C. Suciu, and L. M. Itu, "Applying deep neural networks over homomorphic encrypted medical data," *Comput. Math. Methods Med.*, vol. 2020, 2020, Art. no. 3910250.

[77] A. Kipnis and E. Hibshoosh, "Efficient methods for practical fully homomorphic symmetric-key encrypton, randomization and verification," *IACR Cryptol. EPrint Arch.*, vol. 2012, 2012, Art. no. 637.

[78] S. Yakoubov, "A gentle introduction to Yao's Garbled circuits," preprint on webpage at https://web. mit. edu/sonka89/www/papers/2017ygc. pdf (2017).

[79] A.-R. Sadeghi, T. Schneider, and I. Wehrenberg, "Efficient privacypreserving face recognition," in *Proc. Int. Conf. Inf. Secur. Cryptol.*, 2009, pp. 229–244.

[80] S. Sau, J. Haj-Yahya, M. M. Wong, K. Y. Lam, and A. Chattopadhyay, "Survey of secure processors," in *Proc. Int. Conf. Embedded Comput. Syst., Archit. Model. Simul.*, 2017, pp. 253–260.

[81] H. Dang and E.-C. Chang, "Privacy-preserving data deduplication on trusted processors," in *Proc. IEEE 10th Int. Conf. Cloud Comput.*, 2017, pp. 66–73.

[82] P. R. M. Rao, S. M. Krishna, and A. S. Kumar, "Privacy preservation techniques in big data analytics: A survey," *J. Big Data*, vol. 5, no. 1, 2018, Art. no. 33.

[83] H. Wimmer and L. Powell, "A comparison of the effects of k-anonymity on machine learning algorithms," in *Proc. Conf. Inf. Syst. Appl. Res.*, 2014, Paper 1508.

[84] C. Dwork, "The differential privacy frontier," in *Proc. Theory Cryptogr. Conf.*, 2009, pp. 496–502.

[85] T. Zhu, D. Ye, W. Wang, W. Zhou, and P. S. Yu, "More than privacy: applying differential privacy in key areas of artificial intelligence," *arXiv preprint arXiv:2008.01916*, 2020.

[86] J. Shlens, "A tutorial on principal component analysis," 2014, *arXiv:1404.1100*.

[87] J. Yang, J. Liu, and J. Wu, "Facial image privacy protection based on principal components of adversarial segmented image blocks," *IEEE Access*, vol. 8, pp. 103385–103394, 2020.

[88] M. N. Sadat, M. M. Al Aziz, N. Mohammed, S. Pakhomov, H. Liu, and X. Jiang, "A privacy-preserving distributed filtering framework for nlp artifacts," *BMC Med. Inform. Decis. Making*, vol. 19, no. 1, pp. 1–10, 2019.

[89] H. Wang, S. Gong, X. Zhu, and T. Xiang, "Human-in-the-loop person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 405–422.

[90] I. Rahwan, "Society-in-the-loop: Programming the algorithmic social contract," *Ethics Inf. Technol.*, vol. 20, no. 1, pp. 5–14, 2018.

[91] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *Int. J. Hum. Comput. Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[92] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," 2016, *arXiv:1611.03814*.

[93] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: A survey and review," *arXiv preprint arXiv:1412.7584*, 2014.

[94] "Privacy and data protection by design: From policy to engineering," Eur. Union Agency Network Inf. Security, Heraklion, Greece, 2014.