Privacy, Policy, and Compliance in Yet Another 'Age': The Necessity of Interdisciplinary Collaboration for Artificial Intelligence Applications

Bailey Kacsmar
University of Alberta
kacsmar@ualberta.ca

Abstract

The advent of artificial intelligence (AI) interfaces that are accessible to the general populace, in the form of large language model (LLM) chat bots, brought the field of AI to the forefront of conversations on technology laws. These highly visible LLMs train on massive corpora of data from across the internet, amplifying questions as to the privacy implications of AI and the challenges for transparency, explainability, and auditing. While the recent proliferation of LLMs has increased awareness of the use and misuse of far-reaching technologies, governance and privacy for AI must not ignore past lessons on regulating technologies. With this work, we emphasize the existing domains relevant to AI governance and argue that while there are challenges and nuances for privacy, policy, and compliance in AI, it is generally still automation. Rather than focusing on specific algorithms, data set sizes, or parameter settings, we put forth an organization of the different life-stages that an AI deployment goes through as well as the lessons from relevant sub-fields. We highlight how focusing on the impact and consequences, both intentional and unintentional, provides a better grounding for domain experts and provides the path for connecting technical communities and governance communities to make effective regulations and policies for AI.

1 Introduction

The "Internet Age", "Digital Age", or "Information Age", depending on your terminology of choice, refers to the time period of the mid-twentieth century and is largely associated with the impact of information technology and computers on society [25–27]. There is no clear demarcation between the information age and the recently termed "age of AI" corresponding to the rise in prevalence of artificial intelligence (AI). However, we can largely attribute the prevalence of the phrase, "age of AI", to the significant shift in a specific sub-field AI, that of language modeling, between the years of 2019 and 2023. In particular, the release of a chat interface from OpenAI in 2022 accelerated public facing systems explicitly marketed as AI [97], such as AI chat bots incorporated into search engines [72] and social media platforms [15] as well as chat bot specific platforms [38, 57, 114, 149].

Legal questions and consequences in regards to AI are both currently increasing at a notable rate. For instance, proponent organizations for AI use in high-risk settings, like cognitive therapy, have persisted despite persistent and tragic consequences [73, 89, 103]. Efforts towards adoption in high risk domains have given rise to legal questions and legal actions. For example, there are ambiguities in regards to whether therapy-style exchanges between a person and a chatbot will be held to the same standards of protection requirements that an actual cognitive therapist would be held to. Despite such legal questions, chatbots have already been put forth for use as a therapist [128]. There is also already a record breaking large class action lawsuit for copyright violations against organizations which develop

and release generative AI [10]. Further, AI apps have been displaying user interactions with the chatbot that contained personal information such as both medical and legal topics [124]. Recognition of public concern that has arisen from these instances and others has already influenced regulators in regards to broader data protection regulations and privacy, such as in the case of the UK Commissioner's Office guidance about data collection and data transparency [17].

As public access to tools marketed as 'AI' increased rapidly, concerns about potential harms have also increased, as seen in emergent regulations as well as in media articles and statements from advocacy groups [13, 49, 50, 94, 104, 118]. The EU AI Act, which went into force in 2024, with full effect in 2027, was proposed in 2021 [49]. In mid-2025 a collection of organizations brought forth "The People's AI Action Plan" as a counter view to the Trump administration's stance on AI policy [118]. The people's plan emphasizes that AI should be guided and regulated in ways that center social good, public well-being, accountability, equality, and environmental preservation. Also in 2025, the G7 leaders released a statement on AI in which they set out a human-centered view which advocates for many of the same things of the people's plan, such as considerations for energy demands, impact on majority world countries, as well as impact on the workforce [54, 93]. As AI increases in use, mitigating harms associated with it, specifically bias, surveillance, job loss, and misinformation require greater efforts from appropriate governance [94, 104, 151].

Calls to action, from the public and relevant technical experts, in regards to AI regulation is reminiscent of calls in prior "Ages" where concern for the harms from quick adoption of technology were prevalent [75]. For instance, consider the industrial age and cloth-making. Those most impacted by new machines for spinning and making cloth were the industry professionals in the textile industry. These workers would become known as the Luddites for their objections to the factory model. Their concerns echo the modern day in regards to job loss and injury from insufficient labor protections and lack of safety standards which historically led to loss of life and major injuries for early-adopters of the new technologies [102]. While the abysmal safety standards in the early factories are regarded in modern times with great negativity, the consequences of these practices were not unknowable, even at the time. We, in the current age, do not have to make these mistakes. Instead we can recognize the historical patterns of fast adoption of technology leading to harms, whether it is cloth-making, automobiles, or social media [75, 98, 140]. Specifically, we can initiate the development of what prerequisites are needed to ensure an AI deployment can satisfy existing requirements for each application domain, and support research and innovation efforts towards creating those prerequisites. We are not facing a challenge of how to balance regulation and innovation, a not uncommon juxtaposition [61], rather we have the opportunity for regulation to be used to guide innovations we otherwise would not recognize are needed. In this work, we focus on a path towards identifying such innovations, that may address harms to social and functional norms associated with privacy, including both technical solutions as well as policy and compliance efforts.

Contributions. We emphasize that the literature on AI, governance, privacy, and their intersections is vast. Significant bodies of work exist on protections of privacy in AI, largely focused on machine learning (ML) [53, 117], on AI governance [2, 7, 96], and on communication and social implications of AI and automation [12, 36, 77, 81, 109]. With this paper we do not aim to match the depth in each and every one of these disciplines. Rather, we bring them together in this work to highlight the breadth of these areas and how what is known relates to the field of AI, both in terms of possibilities and limitations. To this end, we provide an overview of relevant subdomains including privacy, policy, compliance, and the field of AI. We then walk through the stages of AI to illustrate where challenges exists, what solutions have been investigated, and the different stakeholders and decisions they make throughout the stages of an AI deployment.

Organization. This paper is organized as follows. Section 2 contains an overview of the concept of privacy and particular interpretations of privacy that intersect with the domain of AI. We discuss the meaning of the terms policy and compliance in Section 3 with a focus on highlighting the differences between how the term 'policy' is employed in computing science and engineering research in discussions of regulatory measures and governance versus how policymakers and related decision makers use the term policy. Section 4 is a review of how the term artificial intelligence has evolved over time, with a focus on how the visibility of AI has changed over time. Finally, in Section 5 we present our discussion of each of the above concepts throughout our structure of AI life stages where we discuss how at each stage decisions are made that impact who an AI application impacts with consequences, where existing technical mitigation strategies can be employed, and how these vectors relate to efforts to regulate AI and evaluate AI applications for compliance.

2 Privacy

Conceptually, the notion of *privacy* is incredibly complex. It is influenced by societal and cultural norms, as well as an individuals own understanding, preferences, and expectations in regards to having privacy. Privacy, while complex, is also very intuitive and simple. While these may seem like contradictions, the reality is that the simplicity corresponds to everyone having an intuitive notion, or mental model, of what is meant when they hear privacy. For instance, we have insight into people's mental models of privacy for hundreds of participants across all ages from Oates et al., where their participants each provided illustrations of what privacy meant to them, some of which were turtles, some locks, and some bathrooms [110]. In terms of capturing the complexity of privacy, there have been many efforts at formalizing privacy including Nissenbaum's *Contextual Integrity* [106, 107], Westin's categories of privacy attitudes [87, 147], and Solove's theory of privacy [132].

If we both generalize and simplify these theories of privacy, then we can takeaway that they each formalize some notion of what is being protected and from who, and who gets to decide what is being protected and from who. In the case of contextual integrity, this formalization is done via norms and the appropriateness of information flows, where both norms and information flows have very particular meanings within the theory of contextual integrity. For Westin's categories of privacy attitudes the formalization is done through illustrating what an individual would be willing to share and with who, based upon what category of individual they are within the three categories of a privacy fundamentalist, privacy pragmatist, or unconcerned. Finally, Solove's taxonomy also captures what is being protected, from who, and who gets to decide by synthesizing different notions of privacy and considering the implications of different settings in terms of privacy. While each of these theories, and more [3, 120], could be applied in depth to understanding and formalizing privacy in the context of AI, we focus on the mitigating and protecting efforts, in particular ones that are reflected in proposed and existing regulations. We discuss different approaches in the context of how they are framed as proxies for privacy and the consequences of such framings. We emphasize that while we put forth each privacy formalization in isolation, specifically data protection, consent, and quantifiable measures, none of these will be sufficient on their own and we highlight weaknesses for each.

2.1 Data Protection as Privacy

Data protection is not necessarily privacy protection. However, there are both regulatory frameworks, such as the European Union's General Data Protection Regulation (GDPR) [142] and Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) [113], as well as privacy enhancing technologies, such as differential privacy [45], that work within a worldview where they prioritize protecting data. The existence of data protection laws does not prevent the existence of additional

privacy laws like in the EU [37]. However, even the existence of privacy focused regulations does not prevent the usage of data protection regulations as a proxy for privacy protections. That is, systems, both technical and regulatory, are formulated around the idea of data. Such protections may encompass what is permissible in terms of how data is collected, how it is used, and how it is accessed. However, in the most strict sense, this abstract notion does not necessarily capture the privacy implications, as it does not have a formal connection to how the use or processing of the data impacts the subject of that data. When we use the term data in computing science, engineering, and even within regulatory documents, the term is a useful and convenient abstraction that captures the notion that there is some form of information that is about something. In practice, that something the information is about, is often a person, who is the subject of the data.

Definition 2.1 (Adapted from Def. 2 [76]) *A data subject is an entity whose data, including information about them or generated by their action or inaction, is present in the data set being computed over (e.g., a training set for ML or statistical analysis) and the data describes the subject or their attributes.*

Definition 2.2 (Adapted from Def. 3 [76]) *A data owner is an entity that holds a dataset that is being contributed to some data analysis (e.g, towards training a ML model or statistical analysis) which is made up of data that originates from one or more data subjects and may or may not include the data owner as one of the data subjects.*

In the above two definitions, neither specifies what the data actually is nor its attributes. For instance, whether the data is considered personally identifiable information (PII), pseudonymous, or anonymous is a factor by some regulations for how it is treated [48, 113, 153]. However, the reality is there is no consensus on a rigorous interpretation of these terms such that there are clear delineations as to whether something is, for example, sufficiently anonymous. Different treatments of data based on how it has been classified by regulations may also fail to capture how it might impact the subjects of the data and whether they agreed for their data to be used in such cases. No matter how the data is protected or perturbed, there is always something that is being learned from that data or the analysis would not be done in the first place [153]. As demonstrated in a 2023 study from Kacsmar et al. [76] the necessity of something being learned does not escape members of the populace, who in general are the data subjects whose information is used in the analysis, as highlighted by a participant who stated "At the end of the day, they're still like learning specific things about me" (P7) [77].

When speaking of data and data analysis, not only is something necessarily being learned, but also there are necessarily one or more parties that contribute to the analysis and correspondingly learn the results of the analysis. Who these parties are, which of them learn the outputs of the analysis, and what type of industry they are in; all impact the perceived acceptability for members of the populace [78]. Despite the impact of these factors, regulatory treatments of data as privacy protection and technical protections of data to protect privacy are formulated and presented using the abstraction of "data". Data is what is being protected, and thus is what is being regulated, despite the issues associated with the abstraction as stated by E.M. Renieris, that "Trying to regulate data as such is like trying to regulate technology as if it has a common definition or clear contours-an exercise in futility" [123]. Certain domains, in particular health data, are sometimes treated as special, with their own laws, beyond the general concept of data [138]. However, despite additional protections being placed on certain types of data in a particular jurisdiction, it is still possible for it to fall outside of the scope of the particular laws. For instance, there are cases of health data falling outside of health focused regulations and leading to cases where datasets with health information are being sold [115]. There are even specific cases where the sold datasets include mental health conditions alongside demographic information of the person or individual names and home regions [33]. Thus, the notion of data protection as privacy is lacking

any human-centered considerations for privacy. One common way of incorporating a small amount of human-centered consideration is via our next notion, using consent management as privacy.

2.2 "Consent" Management as Privacy

The plethora of opt-out, click-to-continue, consent windows that have become the norm in our society are an example of *consent management* as privacy, and an illustration of its weaknesses [9]. Specifically, as regulations required companies to communicate when they collect data, what data they collect, and what they collect it for, these companies sent out updates to privacy policies, added cookie banners to their web pages, and other such modifications as the strategy for being compliant with regulations through a consent-based process [67, 112, 135]. However, while there are a multitude of prompts requesting someone agree before they use a service or platform, it is a stretch to refer to the responses to these prompts as consent. The privacy policies and cookie banners, written and provided by legal experts for an organization, are provided to any person wanting to use the system as an all or nothing agreement. If you want to file your taxes using a particular tax software, you are not provided with a way to do so without agreeing to their policy. In short, these consent management strategies, such as privacy policies can be considered an specific example of the general class termed *contracts of adhesion* [69, 82]. A class of contracts that are hardly read and may even include terms and conditions that should not or cannot be enforced. Privacy policies are a core document used to convey how data is used and collected by companies despite it being long established that these documents are hard to read and rarely read [101, 111]. The inaccessible nature of privacy polices, and the inclusion of contentious terms, is even reflected in a decision in Canadian privacy law. Specifically, in Canada (Privacy Commissioner) v. Facebook, Inc. 2024 FCA 140, which decided Facebook failed to meaningfully get consent from its users, it was stated that:

> "Whether consent is meaningful takes into account all relevant contextual factors; the demographics of the users, the nature of the information, the manner in which the user and the holder of the information interact, whether the contract at issue is a one of adhesion, the clarity and length of the contract and its terms and the nature of the default privacy settings" [21].

The challenges associated with using "consent" as a proxy for privacy do not end at inscrutable and lengthy privacy polices. We also have to account for how organizations have been found to use techniques to manipulate people into selecting options that may not be in their best interest. These techniques are broadly referred to as dark patterns and can be found in online shopping sites, mobile apps, and digital interfaces more broadly [18, 63, 99]. This is not to say consent should not be considered. Rather we mention both the issues with privacy policies and manipulative practices to highlight that while systems "request" people indicate agreement to some practice, these people have other priorities beyond reviewing privacy documents. Instead of expending time deciphering legal jargon, people are focused on whatever primary task is associated with the software they are accessing. For instance, consider the company Wealthsimple which purchased another company, SimpleTax [66]. Shortly after this purchase, the prior promise from Simpletax that indicated they would never sell your data was removed, indicating that going forward continuing to use the software to file taxes in the future would require agreeing to their terms and conditions; which no longer promised not to sell your data. However, an individual facing this update would have to decide to leave an ecosystem that they had previously been using for the important task of completing their taxes. The question then becomes, is a person who continues to file their taxes with this software actually consenting to the new terms and conditions, or are they just trying to do what has to be done, and get their taxes in on time?

It may be tempting to evade this problem by stating that a person who chooses to use a platform despite it lacking privacy protections is choosing to act against their own privacy interests. More extreme arguments may take the form that regardless if a person claims to care about privacy, their actions indicate that they do not hold such values. This framing, that people who act against their own privacy interests, while they claim to care about privacy, is referred to as the *privacy paradox* [43, 85]. However, this is an oversimplification of the reality people face. That is, we must consider the infeasibility of being informed via privacy policies, the use of dark patterns, and that people have a different primary task than preserving their privacy which all together means that while these people may act against their own privacy interests, the reality is that it is incredibly hard for them to do otherwise [122]. The natural followup to this conclusion is that since it is too difficult, essentially infeasible, for the populace to preserve their own privacy interests, technologists and regulators must develop solutions to aid in protecting these individuals privacy. Once such solutions emerge, we next must determine a way to decide whether the solutions are sufficient and a way to test that sufficiency.

2.3 Quantifiable Measures as Privacy

Technologists, when making technical systems, develop *quantifiable measures* to evaluate their work. Technical solutions for privacy have several measures that serve as proxies for how much privacy is lost or the maximum amount of privacy that can be protected. For example, there are technical notions of data anonymity, syntactic notions of privacy, which formalize a particular way of enforcing data anonymity. These notions include k-anonymity [126, 127], ℓ-diversity [95], and t-closeness [90]. Consider the following high-level formulations.

Definition 2.3: For *k-anonymity* [126, 127], there must be at least k records that match any subset of potentially identifying values (quasi-identifiers) that are returned for a query on a given dataset.

Definition 2.4: Extending k-anonymity, *ℓ-diversity* [95] adds the requirement that for any sensitive attribute, there should be at least ℓ distinct values represented in the returned response.

Definition 2.5: Once again extending the data protection notion, the property of *t-closeness [90]* adds the requirement that the distribution of the sensitive attribute and the distribution of the whole data sample should differ by no more than a threshold of t.

These three notions, which provide increasingly formal requirements for data protection, have their uses, in particular for data analysis and data release. However, they do not lend themselves nicely to privacy in AI systems. AI system do not have the same form of data release and thus cannot employ the same protections as query-response based data releases.

In the case of AI the quantifiable notions of privacy employed are either semantic or empirical. The prevalent semantic notion of privacy is *differential privacy* [45, 46], which can be applied to data directly or to the output of a function. Since its formulation in 2006, variations on DP as well as particular ways of applying DP to specific ML algorithms have been developed [53]. For reference, we include the formulation of ϵ-Differential Privacy and discuss conceptually the privacy guarantees that it provides.

Definition 2.6: (ϵ-Differential Privacy [45]). A randomized mechanism $M : \mathcal{D} \mapsto \mathcal{F}$ provides *ϵ-differential privacy* iff for all neighbouring inputs D, $D' \in \mathcal{D}$, ie., differing in one element, and all subsets $F \subseteq \mathcal{F}$,

$$Pr[M(D) \in F] \leq e^{\epsilon} Pr[M(D') \in F], \text{ where the probability space is M's coin tosses.}$$

When used correctly, differential privacy (DP) can effectively protect against leaking information as to the presence or absence of a data point in a calculation. The general idea is that when using DP, an adversary observing a differentially private output of a mechanism is sufficiently unlikely to be able to distinguish between a case where a data element was included in the dataset the mechanism computed over versus a case where that same data element was not included in the dataset.

One way of measuring the success of this protection, outside of its theoretical guarantees, is through an empirical measure of the effectiveness of an attack on a conventional ML model versus one trained using DP [71, 74]. An attack where an adversary aims to determine whether a target data element was included in the training dataset or not, is aiming to execute a membership inference (MI) attack [131]. Thus, through testing the attacks, we can generate a measure of information leakage both by evaluating training time attacks and test-time attacks, the latter of which encompasses the area of inference attacks and can be targeted at various deployed ML models [55, 68, 117]. Inference attacks executed by an adversary may target leaking information about the training data as well as model parameters. Assessing the success of inference attacks does not guarantee the non-existence of more sophisticated attacks where additional information leakage may occur. However, by testing models for their susceptibility to known inference attacks, we can get a measure of the minimum amount of privacy leakage that is occurring [92], as long as awareness is maintained that there may be greater privacy leakage than can be assessed via this evaluation. In summary, while quantifiable measures can give us guidelines to work towards, they are not all encompassing and may be vulnerable to yet undiscovered privacy attacks. We also cannot only focus on the risk of privacy attacks and must also maintain an awareness that although quantifiable measures are useful to us as technologists, when measuring social or human-centered notions like privacy, they will never be a perfect proxy. Thus, our measurements require careful evaluation, as with all the prior privacy notions we discussed, to determine whether their protections are appropriate and sufficient for any given deployment.

> Privacy is a multifaceted concept. Even when focusing on AI, each of the different notions of privacy need to be utilized collectively and in context for any technological deployment to include appropriate consideration and protection of privacy.

3 Policy and Compliance

Policy and compliance are entwined within any effort to regulate technologies. Who contributes to formulating regulations and policies impacts the identifications of potential harms and relevant protections [7]. The breadth of expertise required is a factor for addressing challenges associated with innovations on what are often already relatively novel technologies. To construct clear and meaningful governance for AI, policymakers and technologists need to be able to use common language and definitions for AI, for policy, and for harm [62, 86]. Otherwise, challenges will persist due to how each of these parties may regard the consequences of violating societal norms and expectations like privacy [2].

3.1 The Meaning of Policy

Policy is a broad-reaching non-specific term used to capture a lot of different ways of documenting expectations or requirements for individuals, groups, and governments. Policies can be written within a company about both their internal practices and their external practices for employees and users of the products or services they provide. Schools can write policies about expected behaviors and procedures for students, instructors, and administrators. Finally, policy can also be used to refer to contracts, regulations, and laws, including those from governments.

Research in computing and engineering refer to recommendations for policy and policymakers regularly, but generally neglect to distinguish between the different forms policy instruments can take [60]. Consider, for example, the Government of Canada, which provides explanatory documents on their website for "policies, directives, standards and guidelines" that capture what these terms mean within the Canadian government [59]. The term *policy* in these documents refers to mandatory responsibilities that face internally, meaning they apply to officials and deputy heads, with *directives* guiding how these internal actors are to comply with policy. Less stringent documents include guidelines, which are voluntary, providing advice and recommendations rather than requirements. *Laws and legislation*, and in some jurisdictions also regulations, are what define protections for the citizens of a populace, via placing requirements that apply to organizations and individuals within the legal jurisdiction. Overall, despite how it is used within computing science, the term policy itself, does not necessarily correspond to a regulation or legislation, though it may contribute to the development of such things.

Definition 3.1 ([59]) *"A policy is a set of statements of principles, values, and intent that outlines expectations and provides a basis for consistent decision-making and resource allocation in respect to a specific issue".*

Mandatory policy instruments, which are enforced via laws and regulations, and more voluntary policy instruments can have beneficial outcomes for the populace. Determining which type of policy instrument is most appropriate is likely outside of the expertise of many technologists. However, by communicating to 'policymakers' whether a given recommendation requires everyone partake in order to have the desired outcomes or whether it is still beneficial when only a few organizations partake, is a good starting point.

3.2 Compliance

The fundamental notion of *compliance* is for parties to follow the requirements of pertinent laws. Which laws are pertinent is a question not just of the technology being used or the domain it is being used in, but also a question of territorial reach. While some laws may apply, not just within their physical jurisdictions, but also to citizens of its jurisdiction when they are outside of it, many laws are localized within geographical borders [48, 142]. Data technologies, including internet services and AI, cross these jurisdictional and geographical boundaries, as the use of such technologies and their corresponding impacts on people are not bound within territories. This leads to courts making decisions as to which jurisdictions' laws are to be applied in the case of cross-jurisdictional cases [108].

Compliance requirements give rise to concerns about monetary cost as well as how to ensure systems are auditable in a way that is transparent to the entities that enforce compliance [64, 129]. As new requirements come into effect, companies need to take action, with actions corresponding to changes in processes that can require additional infrastructure, training time of staff, and even just loss of efficiency as employees adapt to the new processes. As a consequence of these, the actual actions a company takes in their effort to be compliant may focus on following the exact letter of the law in a way that most minimizes impact on the company's status-quo. The ways companies act to comply with regulations significantly impacts whether the regulations provide protections, in other words, whether the spirit of the law holds up [9].

As companies enact their practices for emergent technologies or novel applications of existing technologies, such as AI, it can become a competition to establish the best practices and standards that may influence the development of formal laws and oversight. These proactive policies may be very sound, however, focusing on them presumes that the organization is able to set aside its own business goals to produce a beneficial policy that could be applicable more broadly; a perhaps unfair expectation to put on companies' teams [151]. Practices where an organization complies in this way, evading the spirit of

the law, has been termed "avoision" by scholars [79, 137]. Yew et al. have already examined the potential for avoision in the context of the EU AI Act where they identified behaviors which currently, plausibly, and technically, comply with the act while leading to consequences that are contrary to the spirit of the regulation [152]. Therefore, to ensure laws and policy maintain the important outcomes that motivated their construction, we cannot understate the importance of including compliance considerations and challenges across all stages of developing, deploying, and moderating AI.

Technologists use policy to refer to both mandatory and guiding responsibilities for private and public organizations. Divergence in meaning when crossing communities causes issues on both sides, whether it is how policy makers speak of AI or how AI experts speak of policy. These misunderstandings only serve to escalate the issues in the already difficult domain of compliance.

4 Artificial Intelligence

The phrase artificial intelligence itself leads to confusion and misunderstandings about AI technologies' capabilities. Ensuring accurate expectations of what a technology is and is not capable of, fostering risk awareness, requires understanding mental models of the technology and its terminology [44]. The literature for artificial intelligence suggests that how AI technologies are presented, including whether it is referred to as machines, as tools, or as companions influences what human traits or mental capacities are attributed to it by the populace [30].

4.1 Conceptual Terminology Over Time

General notions of automation and stories of thinking machines go back far in history, however, we see the prominent emergence of the term *artificial intelligence* in the late 1950s, some time before the 1956 Dartmouth summer research project [35]. The focus of the 1956 conference was on intelligence and articulating the concept of intelligence such that machines may simulate it. This theme, of understanding intelligence and endeavoring to simulate it, remains prominent in the field of AI to this day. Modern descriptions, such as what one finds in a textbook, refer to the field of AI as being *"...concerned with not just understanding but also building intelligent entities–machines that can compute how to act effectively and safely in a wide variety of novel situations"* [125]. While there are many notions and definitions which are used to refer to the field of AI, what we ultimately see is a split where one camp typically corresponds to *"making machines think like humans"* and the other *"making machines act like humans"* [32]. Representatives of these camps are not limited to particular sub-domains of AI, rather these views may be found across the areas, including natural language processing (NLP), computer vision, and robotics.

Post-2022 is a world where generative AI, and in particular generative AI based chatbots have become prolific and public facing. OpenAI, the organization that released ChatGPT to the world was founded in 2015 with earlier forms of what came to be known as ChatGPT existing as early as 2018 with GPT-1; though it was far from what its successors would become [97]. Even when the advance to GPT-3 occurred in 2022, visibility was still largely limited as it required API access and was primarily only in the awareness of computing science experts and other technologists. In 2022 though, we see the first user interface form of ChatGPT released to the broader populace. The result has been that ChatGPT and the family of models associated with it known as large-language models (LLMs) have in some ways become synonymous with the term AI itself. That is, as reflected on by Karen Hao in regards to her reporting on the area,

> *"While ChatGPT and other so-called large language models or generative AI applications have now taken the limelight, they are but one manifestation of AI, a manifestation that embodies a particular and remarkable narrow view about the way the world is and the way it should be"* [65].

LLMs more broadly, becoming a center for the AI story is entwined with human perceptions of machines and language, which has been around formally since 1955 when John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon proposed a Dartmouth summer research project on AI. In their proposal, they stated that *"An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves"* [100]. Further elaborating on their intentions, they outlined seven aspects they determined as the aspects of the *"artificial intelligence problem"*. Among these aspects, in fact the second one, was the problem of how a computer could be programmed to use language. The justification for this problem proposed that a lot of human thought consists of the manipulation of what could be termed language. Thus, one might expect that the ability to acquire and use language would be a core component to achieving machines that can address *"problems now reserved for humans"*, and be a way to achieve "artificial intelligence". All together, the field of AI has advanced much since its nascent naming. However, while the public view of AI has been overrun with an awareness of LLMs, both LLMs and other AI have a frequent reliance on ML.

4.2 Artificial Intelligence and Machine Learning

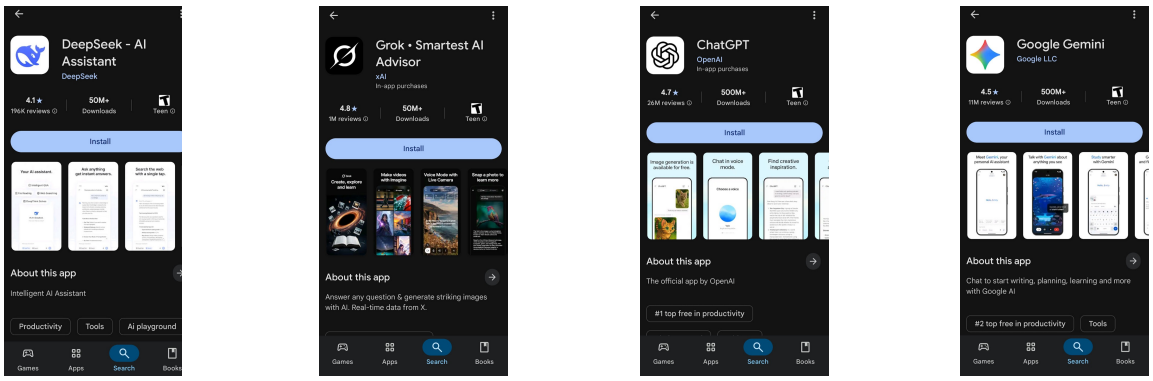
Oftentimes when researchers and technologists say AI what we are talking about is something that uses ML. Modern chat-bots, as well as other modern AI systems, regularly employ ML to achieve the desired functionality. ML is considered a subset of the AI field and the types of ML can be grouped into supervised learning, unsupervised learning, and reinforcement learning [125]. The differences across these types of ML include the goals of the system being produced as well as the way in which data needs to be acquired and prepared for it to be useful to the system. In supervised learning, the goal is to produce a model that is able to perform its task well, such as performing classification, on data it has not seen before. Supervised learning requires training data to be prepared for it with true labels that correspond to the values the model is trying to learn to apply to data. Unsupervised learning takes in training data without labels with a goal of finding patterns that are understandable and of use to humans. Unlike supervised learning, unsupervised learning does not require a corpus of pre-labeled data to train on. Finally, in reinforcement learning the goal is to learn from interacting with an environment such that the system learns a behavior in a way that supports some defined purpose within that environment.

Whether something is AI or ML or even which specific algorithm it uses is not a core issue for this work. However, in the case of policy and compliance there are distinctions associated with the algorithms in use that are of significant importance. For example, different algorithmic approaches, all of which fall under the umbrella of AI, can be either probabilistic or deterministic which has implications for testing, auditing, and transparency. Similarly, model explainability, and whether a model is more or less explainable does not correspond to a particular type of ML. However, some ML algorithms are easier to explain, such as decision trees, whereas others are generally more difficult, such as neural networks [91]. For instance, consider one of the first chat-bots, namely Eliza, which employs rule based systems where outputs are determined via a combination of pattern matching and preset scripts [144, 145]. Possible outputs from Eliza for any given prompt from a user is bounded, as it only outputs things from whatever its current script is set as. While the script can be updated or replaced, ultimately it is possible to reasonably trace why an output was produced. In contrast to this, recent chatbots like DeepSeek [38], Grok [149], ChatGPT [114], and Google Gemini [57] employ large neural networks each of which have

been trained on a large corpus. All together, the amount of data, the parameters and layers of the neural network, along with the different ways the organizations behind these AI models have constructed their output filtration and access controls leads to systems that are infeasible to trace or explain. These chatbots, which are leaders in being hard to explain, are now being deployed en-masse towards having them used in nearly every application one could imagine.

4.3 The Visibility of Artificial Intelligence in Applications

Applications of AI prior to 2022 were largely not marketed as AI to the users of those applications. However, in our post-2022 era, this has completely changed, largely due to the impact of advancements in AI for language, a domain much more accessible, and thus much more marketable to the public than other domains for AI. The goals of language models include both recognition of speech and recognition of text. What is meant by recognition determines a lot. Recognition includes speech-to-text, translation between languages, and predictive text, the last of which is colloquially referred to as auto-complete, an application for AI that has been in wide use, first in web browsers and then in phones, since the early 2000s [56]. AI has also been widely used in recommender systems such as for media consumption on Netflix [4, 14] and YouTube [34], in spam filtration systems [23], in translation [148], in voice assistants like Siri [22], and in fraud detection [8, 105]. Not all AI applications require sophisticated black-box-esque neural networks. Rather, simpler techniques, including linear regression, are also quite effective, including for tasks such as predicting changes in housing markets [19, 24]. Random forest, XGBoost and neural networks, have been employed in efforts for wild fire predictions, further emphasizing that the technical sophistication of the algorithm is not a sole predictor of how useful it is for a task [41]. Even in the case of video games, where there is a greater awareness that AI may be used, the AI includes programmatic solutions in addition to more complex supervised learning, unsupervised learning, and reinforcement learning approaches [150].



(a) DeepSeek (b) Grok (c) ChatGPT (d) Google Gemini

Figure 1: The Google Play store page for the first four results for AI assistant apps on the Google app store as displayed on September 13, 2025.

Notably, the AI in these applications is "hidden" in that it is working in the back-end of the systems. For example the users who access their emails or use their credit cards have no need nor explicit information that the spam folder is produced with the help of AI, or that the flag on their credit card that had a suspicious transaction uses AI. The details of such workings, or even that these systems use

AI at all, stayed with those who made the systems, audited the systems, or otherwise contributed to their production. Those whose primary interactions were based on some other task, such as checking email, did not need nor care to know the details of the automation techniques aiding in their day-to-day interactions. These pre-2022 examples highlight the prolific, but less visible AI of the time.

Post-2022, we see a surge in visibility of AI. Advancement towards producing machines that manipulate language formed through a series of evolutions as to what is the best technique for natural language processing (NLP). These evolutions include, among others, how to process the text by breaking it into parts or n-grams [31], transfer learning [70], and the 'T' in GPT, transformers [141]. After these advancements came forth, we see the transition to AI assistants, AI chatbots, and marketing that uses the term AI. For example, in the Google Play store as of September 13, 2025, there are AI assistants from both long established technology companies as well as more recent ones, which explicitly market their app as "AI", "Intelligent", and "the smartest". We include for reference the app pages of the first four AI assistants shown on our app store in Figure 1. Further demonstrating the extensive way AI has become part of the status-quo in terms of public expectations for technologies, there are even emerging directives from governments on how to adopt AI into their processes [58, 61], a call towards adoption that was not explicit in the regular use of AI predating this point and time.

AI, as with any other technology, cannot be treated as a monolith, nor can it be treated it as a unique innovation that eludes existing regulatory norms on automation. While LLMs have been in the forefront of conversations, AI is a broad field itself with a breadth of applications, each of which have domain-specific norms and standards associated with them throughout any development and deployment.

5 Structuring Life Stages for AI Applications to Identify Privacy Vectors

To determine when and how to use AI, one approach is to consider what the specific technical implementation is, the consequences of its use, and who is impacted by it [96]. In this section we articulate pertinent life stages of AI applications. These stages are identified in part, though consideration for the data science life-cycle. Our stages are formulated to reflect how the processes have changed with advancements of AI as well as to be structured in ways that correspond to policy, regulation, and compliance. Thus our stages deviate from the data science life-cycle. Finally, our stages support the identification of where human actions and decisions are made as well as where technological protections can be implemented when evaluating or assessing novel AI applications.

5.1 Data Science Life-Cycle

The term *data science* broadly refers to an area of analysis which employs statistics, algorithms, and related processes over data sets to extract useful insights. The data science life-cycle as presented by Kelleher and Tierney [80] originates with the CRISP-DM cycle from Chapman et al. [28]. Note there is not just one representation of the cycle, but the CRISP-DM cycle does capture the information relevant to our discussion. The data science cycle allows for moving back and forth between stages, with the stages being defined as 'business understanding', 'data understanding', 'modeling', 'evaluation', and finally 'deployment'. Overall the idea of the data science life cycle is that the likely starting point corresponds to identifying some project objectives and requirements that can be addressed via analysis. Once the project problem is identified, it is now time to collect, clean and prepare, and gain familiarity

with the data for the analysis. The modeling techniques stage could use any type of analysis methods from basic regression through to complex ML techniques depending on the project goal. Once the appropriate modeling technique is identified, it is only a matter of testing and verifying it before finally deploying it to be used for its intended task. We will now use this underlying structure to formulate AI life stages.

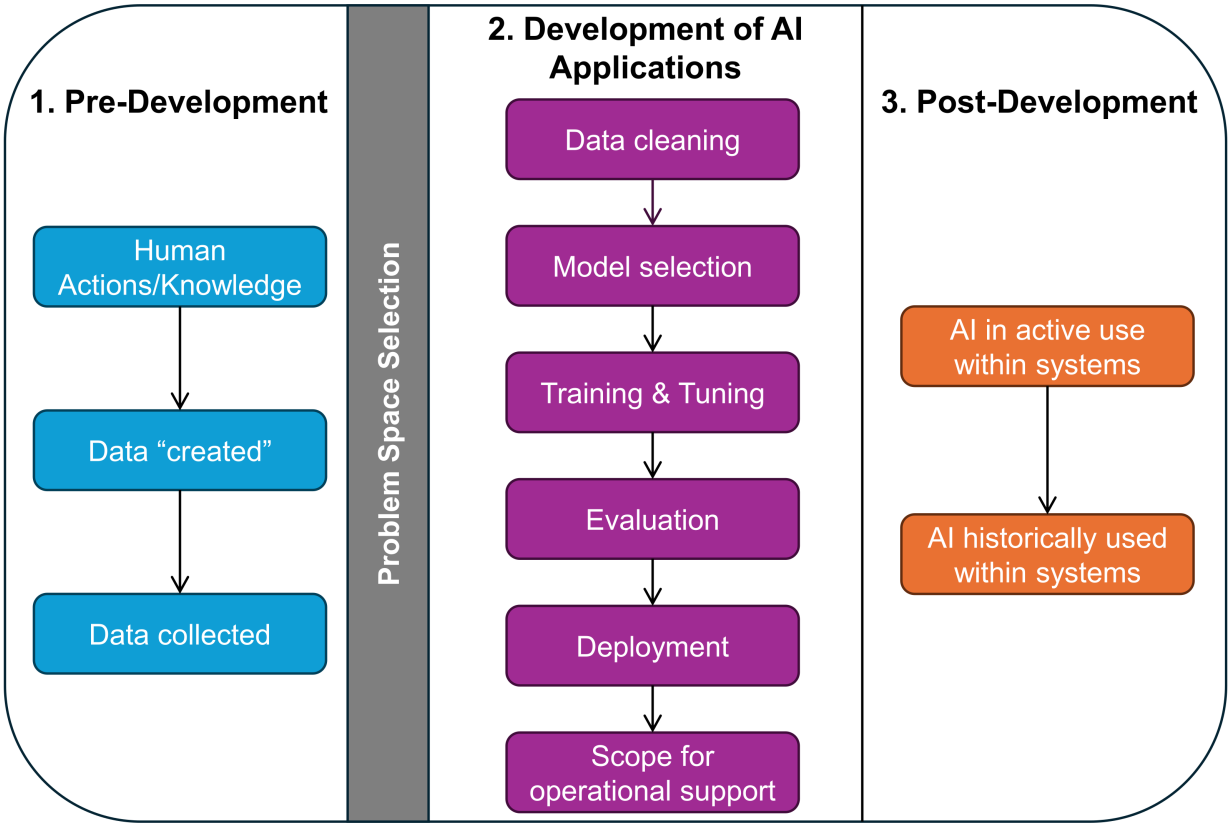


Figure 2: We depict AI application "life" stages of particular relevance for our discussion of privacy, policy, and compliance. These stages span pre-development through to instances where the AI is retired from use. Note that despite our numerical ordering, the process can flow back and forth between stages.

5.2 AI Life Stages

Consider the synthesis of the privacy theories from Section 2 and our claim that the core idea is to define what is being protected, from who, and who gets to decide. Now, by applying this formulation to AI, we can identify the relevant details for each portion of our privacy description. First, in terms of what is being protected, we can formulate it as protecting the training data, protecting the model, protecting the inferences or outputs from the model, and protecting the people impacted by the system at any stage in the AI development process. Second, in terms of who we need to protect against, we may be protecting against the entity that collected all the data (the data owner), those who are the subjects of the data (data subject), individuals within an organization that is involved at any step of the AI process, and even those who just are using the AI system that is produced. Finally, who gets to decide is a reoccurring question across every step of the process. While we can argue that those most

impacted should be the ones that make the decision, the reality is that when these decisions are being made, it can be very far away from those impacted by these decisions. Additionally, those making the decisions, whether company employees or government actors, are powerful and distanced from the people impacted. These decision will also require interdisciplinary expertise to ensure accurate consideration for the functionality of AI as well as the standards within the deployment domain [136]. Through our partitioning of the life stages, we are able to capture the different points in time that people are effected versus when they may actually have any power to intercede on their own behalf or on behalf of others. We provide a visual overview of our AI application life-stages as Figure 2.

5.3 Problem Space Selection

Within our life stages figure, we include *problem space selection* as a literal gray area. Whether the problem space is identified prior to data collection, after data collection, or even after the development of an AI tool, at some point it all comes back to what is the domain and what does it mean to address problems within it. This particular component, while not necessarily a stage itself, is perhaps one of the most critical parts of the AI life-stages. This "stage" is when we can stipulate correctness. For example, before we can measure what "accuracy" is achieved, we need to first ensure that what we have trained the model to assess as being "correct" is actually correct for that application. Furthermore, we may need to account for whether or not the domain has stable or evolving notions of correctness. In the case of fields like medicine or dentistry where practices are continuously advancing or in criminal justice systems where we know historical issues exist, we know that we do not want to mirror the past's standards of correctness [139].

For example, if a classification reflects past ways of referring to someone or something, which are now considered slurs, the model may still probabilistically report that past term since there are a greater number of examples where that was the answer. Reflecting the past probabilistically, whether our social or scientific past, is a risk that may be acceptable for some settings, but certainly not when deploying to any high-impact domains like education or health. In the case of LLMs, they are inherently probabilistic. That is to say, you can give them the same prompt and they could give different outputs. Therefore, LLMs may largely be inappropriate for domains where their probabilistic nature cannot meet the standards of practice for that industry. However, in order to determine that, we cannot make an assessment based on the insights from only technologists, but rather we must engage with domain experts.

5.4 AI Life Stage: Pre-Development

From Daily Life to Data. Throughout any given day, data is generated by our actions and engagements with technology. Every time we access our emails, browse web pages, get captured by video recording doorbells, data is being generated and collected about us. Our day-to-day actions produce immeasurable quantities of data. We define the *pre-development* stage to capture the reality that there are components within the AI life cycle, in particular in relation to data, that may occur well before any consideration or speculation for using AI.

Consider, for example, two "classic" datasets within the field of AI, the Enron dataset and ImageNet [39, 84]. ImageNet came to be through collecting billions of photos from web pages, video clips, and Google's image search database. All of these images had been captured and then shared on the internet by people who had no way of anticipating their images would then be taken in the future, collected together with other images, and labeled by other people to produce a dataset that would then be used to train and evaluate an unknowable multitude of ML models and image analysis algorithms [39]. Similarly, the employees of Enron could not have anticipated their regular email communications would

come to be the Enron dataset, a corpus of text that exists due to a fraud investigation which seized the emails of employees as part of the investigation and then released all the emails to the public. The result was a dataset of communications that has been used by countless researchers in natural language processing [84].

For both the Enron emails and the images within ImageNet, the origin is *human actions or communication of information*. Whether it was taking a photo and sharing it on the internet, sending an email to a spouse, or to a colleague, all of it was captured by researchers and made into datasets. These two datasets, which receive prolific usage, not only did not get consent from the data-subjects, the humans who this data came from, but at the time these people could not even have anticipated their data would be used in this way. Further, we may attempt to argue that such data scraping practices would not be compliant with current regulations. However, the current lawsuit against a company for scraping and utilizing data they gathered from the internet [10] suggests that the practice persists. Data is still being treated as something to be collected for use as a resource to exact value from regardless of the data origins and despite awareness of the financial and personal consequences faced by those whom the data originated with. The extraction and exploitation of our data is so prolific we cannot even begin to understand it with every app we use, every online banking transaction, every credit card purchase, and essentially every action we take connecting back to a digital representation that goes into datasets we know nothing of. Even going so far as data being collected by organizations we have never interacted with such as when data brokers procure data from both public and private sources [20].

From Data to Datasets. After the data was collected, before ImageNet could become what it is today, it first needed to have the mass collection of images be assigned labels, a task that would be relegated to a labor force from across the world which would do the task for low pay through the use of Amazon's Mechanical Turk [5]. The practice of having large scale distributed workforces annotate data quickly and cheaply has become the status-quo in modern day [12, 65, 119]. While the mass collection and annotation of data has become typical, that does not mean overcoming the issues associated with these practices are not the subject of investigation. Attempts have been made to move away from a reliance on large datasets and mitigate the impact they have on people's privacy. For instance, in an effort to go beyond traditional anonymization techniques (recall Section 2.3) the use of synthetic data has been considered. Synthetic data may be generated from sensitive data or from rules and statistics. However, so far synthetic data has not found much success at privacy protection and has not proven better than using traditional anonymization techniques, which themselves are not ideal for training AI [134]. Alternative approaches target the other side of the issue, focusing on transparency and consent, through efforts to formulate how data donation could work [143]. Having a data donation style strategy to address to issues within the pre-development life-stage, while potentially very beneficial, will require a complete overhaul in current practices, re-centering decisions to align with those whom provide the data and those impacted by it.

The people from whom data originates are generally far removed from the decision makers. That is, those impacted are not the same as those who make decisions about data. While data protection laws address some issues, what has become increasingly clear with AI is that the continual treatment of data as something to be possessed or exploited only serves to increase potential harms by distancing these practices from their impact on individuals and society.

5.5 AI Life Stage: Development

The *development* life stage encompasses the aspects of an AI application that are typically determined by technologists, such as algorithmic techniques, parameter configurations, and testing procedures. This includes considering measures for privacy leakage from models via empirical measures of the effectiveness of attacks [71, 74, 92]. However, there are many factors that impact how successful attacks are such as memorization of training data, which is, as one may expect, bad for privacy [133]. However, memorization is not just a bad side effect of the AI training process, but also an important property that corresponds to good performance on data outside of the training sets [51].

This leaves other specialists, who are AI practitioners rather than privacy attack researchers, without clear measures to evaluate against. Despite this, the software developers and engineers are the technical practitioners who code and configure the AI for a system, including the privacy considerations. These considerations can include protecting the data used to train the model as well as protecting against unauthorized access or use of the model. When developing an AI application that requires protecting training data, there are many technical innovations to aid in protecting privacy when training models, including: secure aggregation, differentially private stochastic gradient descent, and differentially private empirical risk minimization among others [1, 16, 29, 42, 53, 116, 130]. However, identifying what the appropriate technique is to use, or even whether there is an appropriate technique is not necessarily within the expertise of the practitioners, and thus there is a need for relevant education, developer tools, and process-oriented support to help practitioners prevent privacy harms being embedded in their deployments [88].

> The technical details, such as the algorithmic techniques and configurations, are determined by technologists within the development stage. However, whether those configurations are appropriate, and how to determine their appropriateness for an application requires insight from outside of the development stage, where domain experts can define what requirements the application has to fulfill and their feasibility.

5.6 AI Life Stage: Post-Development

The *post-development* stage considers the human decisions and actions that impact the consequences and benefits associated with the development of an AI application. After an AI application has been developed for some domain or task, we typically expect it will be deployed. Once the AI system is out in the world, it will either be in a state of active use or it will be retired and withdrawn from use.

First, consider an AI system that is in active use. In this case, maintenance, accounting for changes in what is required from the model, and communicating to those who use the model will require human intervention. Communication is of particular importance, as how people perceive a technology influences their trust in it as much or more than the reality of what the technology can achieve [11, 52, 83, 145]. Mismatched expectations between technological reality and marketing correspond to skewed mental models of AI functionality. Overconfidence in what LLM's can actually do led to a model being used to replace the jobs of people who were staff and volunteers at the National Eating Disorder Association's helpline, at least until the chat bot Tessa had to be removed from use [146]. The chatbot was removed, as it could not actually provide the aid required. This means that people were not acquiring correct mental health resources or appropriate conversational support. Rather, there was an instance where someone who was asking for advice while in eating disorder recovery and the chatbot instead gave information that was essentially suggesting how to continue having an eating disorder. Therefore, the over automation of critical support systems in our society is only leading to greater social and physical

failings for the people who depend on these systems.

Even when an AI application has been retired from use, or removed from use after causing harms like the Tessa chatbot, the impacts of the application can still remain. Consequences from the prior use of an algorithmic deployment, such as an AI application, remain and influence individuals and institutions. This lingering influence is termed an algorithmic imprints by Upol et al. who illustrate its effect in their analysis of an incident of algorithmic deployment on students around the world, focusing on students in Bangladesh [47]. The algorithmic standardization of the results of the General Certificate of Education (GCE) Advanced (A) Level exams in 2020 turned out to be critically flawed and biased, negatively impacting university admissions for students across the world who had taken the exam that year [47, 121]. While the exam grades were ultimately retracted and revised to not use the flawed algorithmic approach, the efforts of the teachers that had to prepare documents and the experience of those students does not go away. Finally, this deployment was inflicted on teachers and students globally by the Office of Qualifications and Exam Regulations in the UK, required significant time investment from the teachers beyond their normal role, and disregarded students preparation efforts. In this particular story, protests and media coverage along with large scale push back eventually corresponded to the retraction of the algorithmic scores, but the power to decide to deploy it and to retract it still remained with the Office of Qualifications and Exam Regulations and not with the students and teachers impacted by it.

There is significant disparity in terms of power as well as pertinent expertise among: those who the data came from, those who decide to use the data for their chosen purpose, those who decide what measure of truth to apply to the data, and those who the resulting AI impacts. Therefore, we must consider whether there should be a way to preserve the right to refuse to participate or be classified by AI.

6 Conclusion

Our understanding of what AI means has changed overtime, as has how we use the term data. The meaning of data changed quickly, but now so too have the consequences of data, such that data protection alone cannot solve the issues of human privacy in our current society. We can make better systems, but even good systems can cause harm without consideration for the full picture of who they impact. The reality is that AI systems necessarily reflect a snapshot in time, the time at which these labels or these notions of correctness were established. Changing these over time requires additional training or additional processes to account for the fact that these models will necessarily not reflect new understandings or new notions of what is acceptable.

Therefore, to advance towards resolving problems at the intersection of AI, privacy, policy, and compliance, we do not need yet another generalized framework or guideline. Even as early as 2021 there were more than 170 guidelines and frameworks in the broad area of responsible and trustworthy AI, including guidelines that synthesized collections of guidelines [6, 40]. Rather, we need to recognize the importance of domain expertise, specifically the expertise required to execute the tasks in the settings where AI applications are being proposed. This means that therapists' expertise should determine the viability of any proposed use of AI in the therapy domain. Dentists should determine the viability of any proposed use of AI in dentistry. Teachers should determine the viability of any proposed use of AI as an educator. This is not to state that these experts should speak on how to implement the AI or whether AI can achieve the standards they identify, but that they are the only ones who can properly determine what a fail state would be for any form of automation working within their area. This is the lesson we can learn from the past ages of industry, we do not need to repeat the mistakes of the past where

"*...workers are time and again ignored by regulators and governments in favor of entrepreneurs and their technologies of disruption*" [102].

To determine what needs to be developed, the technical feasibility of AI, and the appropriate legal consequences of using AI within a particular domain, the fail states must be determine by domain experts. Correspondingly, policy for AI must be guided by domain experts, and when privacy is a factor, privacy experts must contribute as well. We cannot rely on the technology organizations or service providing companies developing AI applications to identify consequences for applications. We must have domain experts collectively work together with law-makers and technologists if we are to develop technology-agnostic requirements on what matters in that domain, what failures cannot happen, and what it means for automation to work correctly.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Vienna, Austria: ACM, 2016, pp. 308–318.

[2] N. Agrawal, R. Binns, M. Van Kleek, K. Laine, and N. Shadbolt, "Exploring Design and Governance Challenges in the Development of Privacy-Preserving Computation," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–13.

[3] I. Altman, *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. ERIC, 1975.

[4] X. Amatriain, "Beyond Data: From User Information to Business Value Through Personalized Recommendations and Consumer Science," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 2201–2208.

[5] Amazon, "Amazon Mechanical Turk," 2025, accessed 2025-11-01.

[6] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen *et al.*, "Guidelines for Human-AI Interaction," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2019.

[7] B. Attard-Frost and K. Lyons, "AI Governance Systems: A Multi-Scale Analysis Framework, Empirical Findings, and Future Directions," *AI and Ethics*, vol. 5, no. 3, pp. 2557–2604, 2025.

[8] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis," in *2017 international conference on computing networking and informatics (ICCNI)*. IEEE, 2017, pp. 1–9.

[9] BBC News, "Cookie Banner Frustration to be Tackled by EU," https://www.bbc.com/news/business-38583001, 2017, accessed 2025-09-18.

[10] A. Belanger, "AI Industry Horrified to Face Largest Copyright Class Action Ever Certified," *Ars Technica*, Aug. 2025. [Online]. Available: https://arstechnica.com/tech-policy/2025/08/ai-industry-horrified-to-face-largest-copyright-classaction-ever-certified/

[11] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models be Too Big? 🦜," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[12] E. M. Bender and A. Hanna, *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. Random House, 2025.

[13] Y. Bengio, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, P. Fox, B. Garfinkel, D. Goldfarb *et al.*, "International AI Safety Report," *arXiv preprint arXiv:2501.17805*, 2025.

[14] J. Bennett and S. Lanning, "The Netflix Prize," *KDD Cup and Workshop 2007*, 2007.

[15] M. Bobrowsky, "Meta Will Begin Using AI Chatbot Conversations to Target Ads," The Wall Street Journal, 2025, accessed 2025-10-20.

[16] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 1175–1191.

[17] R. Booth, "Makers of Air Fryers and Smart Speakers Told to Respect Users' Right to Privacy," *The Guardian*, Jun. 2025. [Online]. Available: https://www.theguardian.com/technology/2025/jun/16/air-fryers-smart-tv-speakers-user-data-privacy-ico

[18] C. Bösch, B. Erb, F. Kargl, H. Kopp, and S. Pfattheicher, "Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 4, pp. 237–254, 2016.

[19] J. Brannlund, H. Y. Lao, M. MacIsaac, and J. Yang, "Predicting Changes in Canadian Housing Markets with Machine Learning," Bank of Canada, Staff Discussion Paper 2023-21, 2023. [Online]. Available: https://www.bankofcanada.ca/wp-content/uploads/2023/09/sdp2023-21.pdf

[20] D. Cameron and D. Mehrotra, "CFPB Quietly Kills Rule to Shield Americans From Data Brokers," Wired, 2025, accessed 2025-05-15.

[21] "Canada (Privacy Commissioner) v. Facebook, Inc., 2024 FCA 140 (CanLII)," https://canlii.ca/t/k6pn1, 2024, acccessed 2025-09-18.

[22] T. Capes, P. Coles, A. Conkie, L. Golipour, A. Hadjitarkhani, Q. Hu, N. Huddleston, M. Hunt, J. Li, M. Neeracher *et al.*, "Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System," in *Interspeech*, 2017, pp. 4011–4015.

[23] G. Caruana and M. Li, "A Survey of Emerging Approaches to Spam Filtering," *ACM computing surveys (CSUR)*, vol. 44, no. 2, pp. 1–27, 2008.

[24] K. E. Case and R. J. Shiller, "Forecasting Prices and Excess Returns in the Housing Market," *Real Estate Economics*, vol. 18, no. 3, pp. 253–273, 1990.

[25] M. Castells, *End of Millennium*. John Wiley & Sons, 2010.

[26] ——, *The Power of Identity*. John Wiley & Sons, 2011, vol. 14.

[27] ——, *The Rise of the Network Society*. John wiley & sons, 2011.

[28] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0," 1999.

[29] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially Private Empirical Risk Minimization," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 1069–1109, 2011.

[30] A. Chen, S. S. Kim, A. Dharmasiri, O. Russakovsky, and J. E. Fan, "Portraying Large Language Models as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–14.

[31] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–394, 1999.

[32] T. Colburn, *Philosophy and Computer Science*. Routledge, 2015.

[33] K. Collier, "A Researcher Tried to Buy Mental Health Data. It Was Surprisingly Easy," *NBC News*, Feb. 2023. [Online]. Available: https://www.nbcnews.com/tech/security/researcher-tried-buy-mental-health-data-was-surprisingly-easy-rcna70071

[34] P. Covington, J. Adams, and E. Sargin, "Deep Neural Networks for YouTube Recommendations," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.

[35] D. Crevier, *AI: The Tumultuous History of the Search for Artificial Intelligence*. Basic Books, Inc., 1993.

[36] R. Cummings, G. Kaptchuk, and E. M. Redmiles, ""I Need a Better Description": An Investigation Into User Expectations For Differential Privacy," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. Association for Computing Machinery, 2021, p. 3037–3052. [Online]. Available: https://doi.org/10.1145/3460120.3485252

[37] N. N. G. de Andrade, "Data Protection, Privacy and Identity: Distinguishing Concepts and Articulating Rights," in *IFIP PrimeLife International Summer School on Privacy and Identity Management for Life*. Springer, 2010, pp. 90–107.

[38] DeepSeek, "DeepSeek," 2025, accessed 13 September 2025. [Online]. Available: https://www.deepseek.com/en

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-Scale Hierarchical Image Database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[40] A. Deshpande and H. Sharp, "Responsible AI Systems: Who are the Stakeholders?" in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 227–236. [Online]. Available: https://doi.org/10.1145/3514094.3534187

[41] F. Di Giuseppe, J. McNorton, A. Lombardi, and F. Wetterhall, "Global Data-Driven Prediction of Fire Activity," *Nature Communications*, vol. 16, no. 1, p. 2918, 2025.

[42] A. Diaa, L. Fenaux, T. Humphries, M. Dietz, F. Ebrahimianghazani, B. Kacsmar, X. Li, N. Lukas, R. A. Mahdavi, S. Oya, E. Amjadian, and F. Kerschbaum, "Fast and Private Inference of Deep Neural Networks by Co-designing Activation Functions," in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 2191–2208. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/diaa

[43] T. Dienlin and S. Trepte, "Is the Privacy Paradox a Relic of the Past? An In-Depth Analysis of Privacy Attitudes and Privacy Behaviors," *European Journal of Social Psychology*, vol. 45, no. 3, pp. 285–297, 2015.

[44] H. J. Do, M. Brachman, C. Dugan, Q. Pan, P. Rai, J. M. Johnson, and R. Thawani, "Evaluating What Others Say: The Effect of Accuracy Assessment in Shaping Mental Models of AI Systems," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–26, 2024.

[45] C. Dwork, "Differential Privacy," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2006, pp. 1–12.

[46] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," *Journal of Privacy and Confidentiality*, vol. 7, no. 3, pp. 17–51, 2016.

[47] U. Ehsan, R. Singh, J. Metcalf, and M. Riedl, "The Algorithmic Imprint," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1305–1317.

[48] European Union, "General Data Protection Regulation (GDPR) - Chapter 1 General Provisions Article 4," https://gdpr-info.eu/chapter-1/, accessed 2025-10-28.

[49] ——, "The EU Artificial Intelligence Act," *European Union*, 2024.

[50] Federal Trade Commission, "FTC Launches Inquiry into AI Chatbots Acting as Companions," https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions, accessed 2025-10-26.

[51] V. Feldman, "Does Learning Require Memorization? A Short Tale about a Long Tail," in *Proceedings of the 52nd annual ACM SIGACT symposium on theory of computing*, 2020, pp. 954–959.

[52] A. Ferrario and M. Loi, "How Explainability Contributes to Trust in AI," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1457–1466. [Online]. Available: https://doi.org/10.1145/3531146.3533202

[53] F. Fioretto, P. Van Hentenryck *et al.*, *Differential Privacy in Artificial Intelligence: From Theory to Practice*. Now Publishers, Inc., 2025.

[54] G7 Leaders, "G7 Leaders' Statement on AI for Prosperity," https://g7.canada.ca/en/news-and-media/news/g7-leaders-statement-on-ai-for-prosperity/, Jun. 2025.

[55] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, October 15-19, 2018*. Toronto, Canada: ACM, 2018, pp. 619–633. [Online]. Available: https://doi.org/10.1145/3243734.3243834

[56] M. Garber, "How Google's Autocomplete Was...Created/Invented/Born," *The Atlantic*, Aug. 2013. [Online]. Available: https://www.theatlantic.com/technology/archive/2013/08/how-googles-autocomplete-was-created-invented-born/278991/

[57] Google, "Gemini," 2025, accessed 13 September 2025. [Online]. Available: https://gemini.google.com/

[58] Government of British Columbia, "Draft Artificial Intelligence Responsible Use Principles," 2024, last updated: November 26, 2024. [Online]. Available: https://digital.gov.bc.ca/ai/draft-responsible-use-principles/

[59] Government of Canada, "Foundation Framework for Treasury Board Policies," 2008, accessed 2025-09-20.

[60] ——, "Introduction to Policy," 2021, accessed 2025-09-20.

[61] ——, "Directive on Automated Decision-Making," 2025, last updated: July 24, 2025. [Online]. Available: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

[62] N. Guha, C. M. Lawrence, L. A. Gailmard, K. T. Rodolfa, F. Surani, R. Bommasani, I. D. Raji, M.-F. Cuéllar, C. Honigsberg, P. Liang *et al.*, "AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing," *Geo. Wash. L. Rev.*, vol. 92, p. 1473, 2024.

[63] J. Gunawan, A. Pradeep, D. Choffnes, W. Hartzog, and C. Wilson, "A Comparative Study of Dark Patterns Across Web and Mobile Modalities," in *Proceedings of the ACM on Human-Computer Interaction, Volume 5, Issue CSCW2*, vol. 5, no. CSCW2. New York, NY, USA: Association for Computing Machinery, Oct. 2021. [Online]. Available: https://doi.org/10.1145/3479521

[64] P. Hacker, "Comments on the Final Trilogue Version of the AI Act," *Available at SSRN 4757603*, 2024.

[65] K. Hao, *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. Penguin Random House, 2025.

[66] A. Heydari, "The Canadian Tech Company that Changed its Mind About Using Your Tax Return to Sell Stuff," CBC Radio, 2020, accessed 2025-09-18.

[67] M. Hils, D. W. Woods, and R. Böhme, "Measuring the Emergence of Consent Management on the Web," in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 317–332.

[68] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning," in *the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 603–618.

[69] D. A. Hoffman, "Defeating the Empire of Forms," *Virginia Law Review*, vol. 109, no. 7, pp. 1367–1427, 2023.

[70] J. Howard and S. Ruder, "Universal Language Model Fine-Tuning for Text Classification," *arXiv preprint arXiv:1801.06146*, 2018.

[71] T. Humphries, S. Oya, L. Tulloch, M. Rafuse, I. Goldberg, U. Hengartner, and F. Kerschbaum, "Investigating Membership Inference Attacks Under Data Dependencies," in *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. IEEE, 2023, pp. 473–488.

[72] L. Jamali, "AI Chatbot to be Embedded in Google Search," BBC, 2025, accessed 2025-10-20.

[73] J. Jargon and S. Kessler, "A Troubled Man, His Chatbot and a Murder-Suicide in Old Greenwich," https://www.wsj.com/tech/ai/chatgpt-ai-stein-erik-soelberg-murder-suicide-6b67dbfb, accessed 2025-10-26.

[74] B. Jayaraman and D. Evans, "Evaluating Differentially Private Machine Learning in Practice," in *the 28th USENIX Security Symposium*, Santa Clara, CA, 2019, pp. 1895–1912.

[75] S. E. Jones, *Against Technology: From the Luddites to Neo-Luddism*. Routledge, 2013.

[76] B. Kacsmar, "Perceptions and Practicalities for Private Machine Learning," Ph.D. dissertation, University of Waterloo, 2023.

[77] B. Kacsmar, V. Duddu, K. Tilbury, B. Ur, and F. Kerschbaum, "Comprehension from Chaos: Towards Informed Consent for Private Computation," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 210–224. [Online]. Available: https://doi.org/10.1145/3576915.3623152

[78] B. Kacsmar, K. Tilbury, M. Mazmudar, and F. Kerschbaum, "Caring about Sharing: User Perceptions of Multiparty Data Sharing," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 899–916. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/kacsmar

[79] L. Katz, *Ill-Gotten Gains: Evasion, Blackmail, Fraud, and Kindred Puzzles of the Law*. University of Chicago Press, 1996.

[80] J. D. Kelleher and B. Tierney, *Data Science*. MIT press, 2018.

[81] P. G. Kelley, C. Cornejo, L. Hayes, E. S. Jin, A. Sedley, K. Thomas, Y. Yang, and A. Woodruff, "'There will be less privacy, of course': How and why people in 10 countries expect {AI} will affect privacy in the future," in *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, 2023, pp. 579–603.

[82] F. Kessler, "Contracts of Adhesion–Some Thoughts About Freedom of Contract," *Columbia Law Review*, vol. 43, no. 5, pp. 629–642, 1943.

[83] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, "Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 77–88.

[84] B. Klimt and Y. Yang, "Introducing the Enron Corpus," in *CEAS*, vol. 45, 2004, pp. 92–96.

[85] S. Kokolakis, "Privacy Attitudes and Privacy Behaviour: A Review of Current Research on the Privacy Paradox Phenomenon," *Computers & security*, vol. 64, pp. 122–134, 2017.

[86] P. M. Krafft, M. Young, M. Katell, K. Huang, and G. Bugingo, "Defining AI in Policy versus Practice," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 72–78. [Online]. Available: https://doi.org/10.1145/3375627.3375835

[87] P. Kumaraguru and L. F. Cranor, *Privacy Indexes: A Survey of Westin's Studies*. Carnegie Mellon University, School of Computer Science, Institute for ..., 2005.

[88] H.-P. H. Lee, L. Gao, S. Yang, J. Forlizzi, and S. Das, "" I Don't Know If We're Doing Good. I Don't Know If We're Doing Bad": Investigating How Practitioners Scope, Motivate, and Conduct Privacy Work When Developing {AI} Products," in *USENIX Security Symposium*, 2024.

[89] M. Li, W. Bickersteth, N. Tang, L. Cranor, J. Hong, H. Shen, and H. Heidari, "A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 8, no. 2, 2025, pp. 1561–1573.

[90] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy Beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*. IEEE, 2006, pp. 106–115.

[91] Z. C. Lipton, "The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[92] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, "{ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4525–4542.

[93] M. Longiaru, W. Negrón, B. J. Chen, A. Nguyen, S. N. Patel, and D. Calacci, "The 'Privacy' Trap: How "Privacy-Preserving AI Techniques" Mask the New Worker Surveillance and Datafication," https://datasociety.net/library/the-privacy-trap/, accessed 2025-10-26.

[94] L. MacCleery, "The New Surveillance State: Why Data Privacy Is Now Essential to Democracy," *Tech Policy Press*, Jun. 2025. [Online]. Available: https://techpolicy.press/the-new-surveillance-state-why-data-privacy-is-now-essential-to-democracy/

[95] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy Beyond k-anonymity," *Acm transactions on knowledge discovery from data (tkdd)*, vol. 1, no. 1, pp. 3–es, 2007.

[96] M. Mäntymäki, M. Minkkinen, T. Birkstedt, and M. Viljanen, "Defining Organizational AI Governance," *AI and Ethics*, vol. 2, no. 4, pp. 603–609, 2022.

[97] B. Marr, "A Short History Of ChatGPT: How We Got To Where We Are Today," Forbes, 2023, https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/.

[98] J. L. Mashaw and D. L. Harfst, *The Struggle for Auto Safety*. Harvard University Press Cambridge, MA, 1990.

[99] A. Mathur, G. Acar, M. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan, "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites," *Proceedings of the ACM Human-Computer Interaction*, vol. 1, 2019.

[100] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI magazine*, vol. 27, no. 4, pp. 12–12, 2006.

[101] A. M. McDonald and L. F. Cranor, "The Cost of Reading Privacy Policies," *ISJLP*, vol. 4, p. 543, 2008.

[102] B. Merchant, *Blood in the Machine: The Origins of the Rebellion Against Big Tech*. Hachette UK, 2023.

[103] B. Montgomery, "Mother Says AI Chatbot Led her Son to Kill Himself in Lawsuit Against its Maker," https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death, accessed 2025-10-26.

[104] Nature Editorial, "Stop Talking About Tomorrow's AI Doomsday When AI Poses Risks Today," *Nature*, vol. 618, pp. 885–886, 2023.

[105] E. W. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and An Academic Review of Literature," *Decision support systems*, vol. 50, no. 3, pp. 559–569, 2011.

[106] H. Nissenbaum, "Privacy in Context: Technology, Policy, and the Integrity of Social Life," in *Privacy in context*. Stanford University Press, 2009.

[107] ——, "Contextual Integrity Up and Down the Data Food Chain," *Theoretical Inquiries in Law*, vol. 20, no. 1, pp. 221–256, 2019.

[108] M. Nitzberg and J. Zysman, "Algorithms, Data, and Platforms: The Diverse Challenges of Governing AI," *Journal of European Public Policy*, vol. 29, no. 11, pp. 1753–1778, 2022.

[109] S. U. Noble, "Algorithms of Oppression: How Search Engines Reinforce Racism," in *Algorithms of oppression*. New York university press, 2018.

[110] M. Oates, Y. Ahmadullah, A. Marsh, C. Swoopes, S. Zhang, R. Balebako, and L. F. Cranor, "Turtles, Locks, and Bathrooms: Understanding Mental Models of Privacy Through Illustration," *Proceedings on Privacy Enhancing Technologies*, 2018.

[111] J. A. Obar and A. Oeldorf-Hirsch, "The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services," *Information, Communication & Society*, pp. 1–20, 2018.

[112] S. O'Connor, R. Nurwono, A. Siebel, and E. Birrell, "(Un)clear and (In)conspicuous: The Right to Opt-out of Sale Under CCPA," in *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*. New York, New York: ACM, 2021, pp. 59–72.

[113] Office of the Privacy Commissioner of Canada, "PIPEDA in Brief," https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/, 2019, accessed 2019-06-18.

[114] OpenAI, "ChatGPT," 2025, accessed 13 September 2025. [Online]. Available: https://chatgpt.com/

[115] D. Otis, "Private Clinics in Canada are Selling Personal Health Data: Study," *CTV News*, May 2025. [Online]. Available: https://www.ctvnews.ca/health/article/private-clinics-in-canada-are-selling-personal-health-data-study/

[116] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-Supervised Knowledge Transfer for Deep Learning from Private Training Data," in *the International Conference on Learning Representations*, Toulon, France, 2017.

[117] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and Privacy in Machine Learning," in *the 2018 IEEE European Symposium on Security and Privacy*. London, UK: IEEE, 2018, pp. 399–414.

[118] "People's AI Action Plan," https://peoplesaiaction.com/, 2025, accessed on September 10, 2025.

[119] B. Perrigo, "Is 'Sweatshop Data' Really Over?" *TIME*, Jul. 2025. [Online]. Available: https://time.com/7306153/ai-sweatshop-data-over/

[120] S. Petronio, *Boundaries of Privacy: Dialectics of Disclosure*. Suny Press, 2002.

[121] M. Posner, "The Fight to Get AI Right in Bangladesh, the World's Eighth-Most Populous Country," *Khoury News*, Jul. 2025. [Online]. Available: https://www.khoury.northeastern.edu/the-fight-to-get-ai-right-in-bangladesh-the-worlds-eighth-most-populous-country/

[122] K. Renaud, M. Volkamer, and A. Renkema-Padmos, "Why doesn't Jane protect her privacy?" in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2014, pp. 244–262.

[123] E. M. Renieris, *Beyond Data: Reclaiming Human Rights at the Dawn of the Metaverse*. MIT Press, 2023.

[124] K. Robison, "The Meta AI App Lets You 'Discover' People's Bizarrely Personal Chats," *WIRED*, Jun. 2025. [Online]. Available: https://www.wired.com/story/meta-artificial-intelligence-chatbot-conversations/

[125] S. Russell, P. Norvig, and A. Intelligence, "Artificial Intelligence: A Modern Approach," *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, no. 27, pp. 79–80, 1995.

[126] P. Samarati, "Protecting Respondents Identities in Microdata Release," *IEEE transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2002.

[127] P. Samarati and L. Sweeney, "Protecting Privacy When Disclosing Information: k-anonymity and its Enforcement Through Generalization and Suppression," technical report, SRI International, 1998.

[128] R. Scammell, "Sam Altman Says Your ChatGPT Therapy Session Might Not Stay Private in a Lawsuit," *Business Insider*, Jul. 2025. [Online]. Available: https://www.businessinsider.com/chatgpt-privacy-therapy-sam-altman-openai-lawsuit-2025-7

[129] M. Shafieinejad, X. He, and B. Kacsmar, "Adopt a PET! An Exploration of PETs, Policy, and Practicalities for Industry in Canada," *arXiv preprint arXiv:2503.03027*, 2025.

[130] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in *the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1310–1321.

[131] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *the 2017 IEEE Symposium on Security and Privacy*. San Jose, CA, USA: IEEE, 2017, pp. 3–18.

[132] D. J. Solove, *Understanding Privacy*. Harvard university press, 2010.

[133] C. Song, T. Ristenpart, and V. Shmatikov, "Machine Learning Models that Remember Too Much," in *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, 2017, pp. 587–601.

[134] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic Data–Anonymisation Groundhog Day," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1451–1468.

[135] State of California Department of Justice, "California Consumer Privacy Act," https://oag.ca.gov/privacy/ccpa, 2018, accessed 2022-09-04.

[136] K. Tilbury, B. Kacsmar, and J. Hoey, "Towards Safety in Multi-agent Reinforcement Learning through Security and Privacy by Design," in *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*, 2024.

[137] R. Turner, "Reactions of the Regulated: A Federal Labor Law Example," *Lab. Law.*, vol. 17, p. 479, 2001.

[138] US Department of Health and Human Services, "Health Information Privacy," https://www.hhs.gov/hipaa/index.html, accessed 2025-10-25.

[139] S. Vallor, *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press, 2024.

[140] J. Van Dijck, *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press, 2013.

[141] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in neural information processing systems*, vol. 30, 2017.

[142] P. Voigt and A. Von dem Bussche, "The EU General Data Protection Regulation (GDPR)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[143] R. Wang, R. De Viti, A. Dubey, and E. M. Redmiles, "The Role of Privacy Guarantees in Voluntary Donation of Private Data for Altruistic Goals," *arXiv e-prints*, pp. arXiv–2407, 2024.

[144] J. Weizenbaum, "ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[145] ——, *Computer Power and Human Reason: From Judgment to Calculation*. WH Freeman & Co, 1976.

[146] C. Westfall, "Non-Profit Helpline Shifts To Chatbots, Then Shuts Down Rogue AI," Forbes, 2023, accessed 2025-07-05.

[147] A. F. Westin, *Privacy and Freedom*. IG Publishing, 1967.

[148] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation," *arXiv preprint arXiv:1609.08144*, 2016.

[149] xAI, "Grok," 2025, accessed 13 September 2025. [Online]. Available: https://x.ai/grok

[150] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer, 2018, vol. 2.

[151] R.-J. Yew and B. Judge, "Anti-Regulatory AI: How 'AI Safety' is Leveraged Against Regulatory Oversight," *arXiv preprint arXiv:2509.22872*, 2025.

[152] R.-J. Yew, B. Marino, and S. Venkatasubramanian, "Red Teaming AI Policy: A Taxonomy of Avoision and the EU AI Act," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025, pp. 404–415.

[153] R.-J. Yew, L. Qin, and S. Venkatasubramanian, "You Still See Me: How Data Protection Supports the Architecture of AI Surveillance," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 1709–1722.