AI Risk Management Should Understand and Account for Both Safety and Security

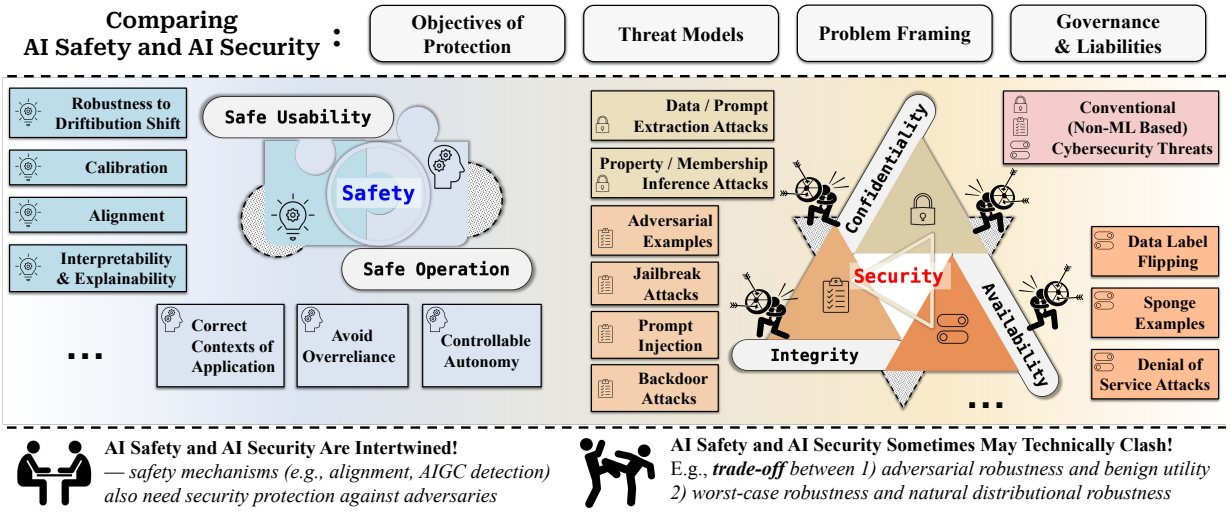
A Preprint

Xiangyu Qi[1], **Yangsibo Huang**[1], **Yi Zeng**[2], **Edoardo Debenedetti**[3], **Jonas Geiping**[4,5], **Luxi He**[1], **Kaixuan Huang**[1], **Udari Madhushani**[6,7], **Vikash Sehwag**[8], **Weijia Shi**[9], **Boyi Wei**[1], **Tinghao Xie**[1], **Danqi Chen**[1], **Pin-Yu Chen**[10], **Jeffrey Ding**[11], **Ruoxi Jia**[2], **Jiaqi Ma**[12], **Arvind Narayanan**[1], **Weijie J. Su**[13], **Mengdi Wang**[1], **Chaowei Xiao**[14,15], **Bo Li**[12,16], **Dawn Song**[17], **Peter Henderson**[1], **Prateek Mittal**[1]

[1]Princeton University [2]Virginia Tech [3]ETH Zurich [4]ELLIS Institute Tübingen [5]MPI for Intelligent Systems [6]Stanford University [7]JPMorgan AI Research [8]Sony AI [9]University of Washington [10]IBM Research [11]George Washington University [12]University of Illinois at Urbana-Champaign [13]University of Pennsylvania [14]University of Wisconsin, Madison [15]NVIDIA [16]University of Chicago [17]UC Berkeley

Abstract

The exposure of **security** vulnerabilities in **safety**-aligned language models, e.g., susceptibility to adversarial attacks, has shed light on the intricate interplay between AI safety and AI security. Although the two disciplines now come together under the overarching goal of AI risk management, they have historically evolved separately, giving rise to differing perspectives. Therefore, in this paper, we advocate that stakeholders in AI risk management should be aware of the nuances, synergies, and interplay between safety and security, and unambiguously take into account the perspectives of both disciplines in order to devise mostly effective and holistic risk mitigation approaches. Unfortunately, this vision is often obfuscated by the poor definitions of the concepts of "safety" and "security" themselves, as their meanings vary considerably across communities. With AI risk management being increasingly cross-disciplinary, this issue is particularly salient. In light of this conceptual challenge, we introduce a unified reference framework to clarify the differences and interplay between AI safety and AI security, aiming to facilitate a shared understanding and effective collaboration across communities.



Correspondence to: Xiangyu Qi (xiangyuqi@princeton.edu), Yangsibo Huang (yangsibo@princeton.edu), Peter Henderson (peter.henderson@princeton.edu), Prateek Mittal (pmittal@princeton.edu)

1 Introduction

The rapid advancement and widespread deployment of the latest foundation models, especially large language models (LLMs) (Brown et al., 2020; Bommasani et al., 2021; OpenAI, 2022, 2023b,c; Touvron et al., 2023a,b; Anthropic, 2023; Gemini Team, 2023; Bubeck et al., 2023), also bring an escalation of AI risks. Enhanced AI capabilities raise the stakes of potential misuse, while the expanded applicability of AI widens the range of possible harm in cases of AI failures. Therefore, stakeholders across disciplines now regard AI risk management as a top priority (Leike & Sutskever, 2023; OpenAI, 2023d; Bengio et al., 2023; European Commission, 2023; Biden, 2023; GOV.UK, 2023).

Currently, a pressing need in AI risk management is to establish a common ground that can effectively coordinate relevant efforts from different disciplines. This issue is salient as the practices of AI risk management are increasingly cross-disciplinary, bringing together practitioners from diverse communities with differing perspectives and focal points in their considerations. This paper sheds light on two such disciplines that are particularly relevant: **AI Safety** and **AI Security**. The boundary between the two is increasingly vague, but they should not be conflated.

Historically, safety and security have long been developed as separate disciplines across a broad range of contexts (Anderson, 2022) and established sectors, such as aviation, nuclear energy, chemistry, power grids, and information systems (Pièrre-Cambacédès & Chaudet, 2010). It has deep roots in AI research as well, as manifested by the separate development of AI safety and AI security studies within the academic literature.[1] Specifically, AI safety initially emerged from concerns about the long-term implications and risks associated with superintelligence or AGI (Yampolskiy, 2013; Bostrom, 2014). Later, Amodei et al. (2016) and Hendrycks et al. (2021b) grounded AI safety within the more specific context of machine learning (ML), identifying hazards that could arise from the adoption of ML systems. In contrast, AI security has largely focused on adversarial threats to AI systems. Notable early studies include adversarial attacks on ML-based spam filters (Dalvi et al., 2004; Lowd & Meek, 2005) and subsequent work that formalized the notion of security in the context of ML models (Venkataraman et al., 2008; Barreno et al., 2010). More recently, adversarial machine learning (AdvML) (Huang et al., 2011; Kurakin et al., 2016; Chen & Hsieh, 2022; Chen & Das, 2023; Vassilev et al., 2024) accounts for a large portion of AI security research.

We advocate that stakeholders in AI risk management should be aware of the nuances, synergies, and interplay between safety and security, and unambiguously take into account the perspectives of both disciplines in order to devise more effective and holistic risk mitigation approaches. A concrete example is that recent alignment approaches (Ouyang et al., 2022; Bai et al., 2022b; Ganguli et al., 2022; Leike & Sutskever, 2023; Inan et al., 2023), predominantly advocated by industry stakeholders for addressing AI safety issues, largely fall short of securing the AI models against adversarial attacks (Abdelnabi et al., 2023; Qi et al., 2023a,c; Carlini et al., 2023b; Zou et al., 2023b; Zou, 2023). Stakeholders must be aware of and transparently communicate the two different sets of considerations, ensuring holistic coverage of investment in addressing both safety and security problems in AI systems.

Unfortunately, this is a position that is simple to make but often severely obfuscated, as the definitions of the basic concepts of "safety" and "security" themselves are often inconsistent and lack consensus. The key conceptual challenge lies in the fact that practitioners from different communities tend to define the two concepts based on their own methodologies and needs, which are seldom consistent. For example, Amodei et al. (2016) frame the discussion of AI safety more conservatively, with a bias towards the problem of accidents in machine learning (ML) systems, while acknowledging security as a closely related field. But Hendrycks et al. (2021b) more directly position many AI security problems under the considerations of AI safety. Interestingly, there are also definitions from the security community that consider AI security as a superset of AI safety instead (Thacker, 2023).

The AI Risk Management Framework (NIST, 2023) recently developed by the U.S. National Institute of Standards and Technology (NIST) instead defines safety and security as two separate characteristics of trustworthy AI. It suggests that AI safety should take cues from and align with efforts, guidelines, and standards for safety in other established sectors, such as transportation and healthcare; while NIST Cybersecurity Framework (NIST, 2018) is recommended as a reference for AI security. This further invites more diverse opinions from all those related sectors.

Essentially, behind safety and security lies the differing mindsets, methodologies, policy and legal considerations, and cultural influences that have been deeply tied to each concept and made profound impacts across a wide range of sectors. Instead of engaging in a futile war on an elusive quest for a universal definition of safety and security (Burns et al., 1992), we propose to establish at least a middle-ground reference framework that can fulfill the basic function of comprehensively illustrating the most differing considerations that underpin safety and security, while also providing the flexibility in modeling their interactions and overlaps. This approach intends to enable

[1] This is also visible in community events. The *ICLR 2021 Workshop on Security and Safety in Machine Learning Systems* (https://aisecure-workshop.github.io/aml-iclr2021/) explicitly noted the differing communities and perspectives between safety and security and called for cross-community collaborations.

interdisciplinary practitioners to find common ground, learn from one another's experiences, and collaborate more effectively toward the shared overarching goal of managing AI risks.

In this paper, *we present such a reference framework in Section 2.* We first elaborate on **distinct objectives of protection** (Section 2.1) that safety and security represent. Safety focuses on preventing harm that a system might inflict upon the external environment and is thus directly confronted with the eventual damages that AI systems may cause to our society, which is more sociotechnical in nature (Lazar & Nelson, 2023). On the other hand, security is more oriented toward protecting the system itself against harm and exploitation from malicious external actors. This characterization of objectives is compatible with recent definitions of AI safety and AI security made by the UK AI Safety Institute[2]. Next, in Section 2.2, we discuss the two **different threat models** that are important in safety and security considerations. In general, security predominantly focuses on the adversarial threat model with the existence of attackers. Safety, in comparison, has a large focus on addressing accidental/unintended harms — meanwhile, adversarial threats are also becoming an increasingly important part of considerations to achieve safety objectives. Based on the differing objectives and threat models, we also discuss how safety and security considerations could differ in their **problem framing** (Section 2.3), **governance and liability** (Section 2.4).

In Section 3, we further elaborate on two distinct taxonomies that encapsulate two differing sets of problems present in the academic literature. The first taxonomy (Section 3.1) presents representative problems that have been predominantly investigated by the AI safety community. The second taxonomy, expounded in Section 3.2, is instead characterized by an AI security perspective. The dual taxonomies offer concrete illustrations of the differing mindsets and considerations underpinning AI safety and AI security.

Based on the nuanced understanding of the safety and security perspectives we have established, in Section 4, we analyze how both perspectives have critical implications in frontier problems that AI risk management is faced with. We call for bridging AI safety and AI security in practice while also discussing how the two can sometimes technically clash. In Section 5, we conclude.

2 A Reference Framework for Comparing Safety and Security in AI Systems

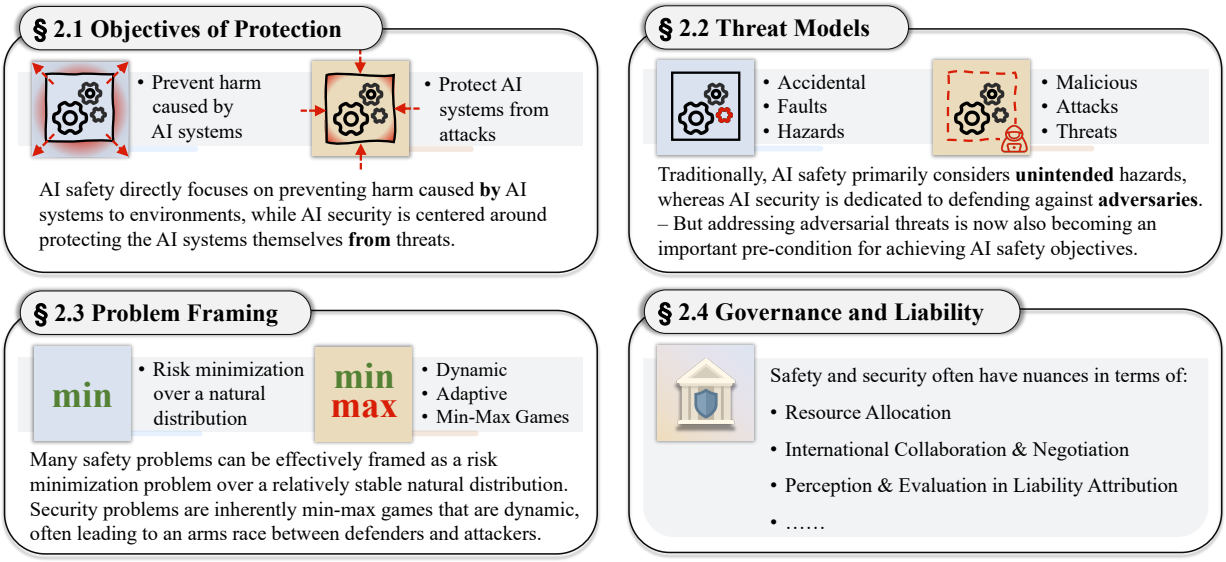


Figure 1: We present a reference framework to systematically examine the differing considerations that underpin AI safety and AI security, and discuss their interplay. We elaborate on four dimensions: objectives of protection (Section 2.1), threat models (Section 2.2), problem framing (Section 2.3), and governance and liability (Section 2.4).

2.1 Objectives of Protection

Safety and security, in general, represent two different objectives of protection. Safety focuses on *preventing harm that a system might inflict upon the external environment.* Security is oriented toward *protecting the system itself*

[2] https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute#definitions

against harm and exploitation from malicious external actors. Contextualizing this in AI risk management, safety considerations focus on addressing undesirable consequences following AI failures. This could include near-term issues such as self-driving cars hitting pedestrians (Kohli & Chadha, 2019); AI models deployed in socially critical contexts generating unintended toxic content (Gehman et al., 2020), perpetuating bias (Abid et al., 2021), and producing misleading information (Kreps et al., 2022); as well as long-term concerns like uncontrolled rogue AIs (Hendrycks et al., 2023; Carlsmith, 2022). In contrast, the security considerations emphasize protecting specific security-critical components, properties, or functionalities of AI systems. For example, the integrity of training data against data poisoning (Chen et al., 2017; Goldblum et al., 2022), the integrity of model inference against adversarial examples (Goodfellow et al., 2014), the confidentiality of sensitive training data against inference attacks (Shokri et al., 2017) and extraction attacks (Carlini et al., 2021; Nasr et al., 2023a).

This distinction could shed light on a recent debate on whether jailbreaking attacks (Zou et al., 2023b; Qi et al., 2023c) against aligned LLMs should be of concern. Critics often cite the example of *asking the model about building a bomb*, arguing that responses from jailbroken models, while inappropriate, are naive and impractical, thus not posing practical safety hazards. However, a security perspective could instead argue that the guardrail is a critical property of the system, and the fact that adversarial attacks can break this property itself is a security failure. This perspective suggests that the attacks and defenses surrounding these guardrails constitute a self-contained security game, even though the subsequent safety hazards that come from current jailbroken models to the external environment are moderate. By focusing particularly on the abstract integrity property of the guardrail itself, a decoupled security study prepares us for better management and control over increasingly powerful models, from which the potential safety hazards are likely to only keep escalating. Such abstraction is a very fundamental idea in security thinking. Back in 100 BC, Julius Caesar used cryptography primarily to convey secret messages to his army generals (Kahn, 1996), but the underlying security principles are now applied to protect the confidentiality of everything, from our daily private messages to critical infrastructures such as nuclear weapons. Dedicated investment in the study of critical AI security principles may similarly prove rewarding in the long run.

We also note that achieving safety and security objectives often involves differing expertise and techniques guided by differing methodologies. AI safety is directly confronted with the eventual damages that AI systems may cause to our society and would thus be more sociotechnical in nature (Lazar & Nelson, 2023). AI security instead revolves around protecting particular AI systems, requiring proficiency in adversarial thinking, such as threat modeling of various possible malicious attacks and implementing robust defensive techniques against adaptive adversaries.

The two different objectives of safety and security also have casual interplay (Burns et al., 1992). Specifically, when some security-critical properties of an AI system are compromised, it can (but not necessarily) further lead to abnormal system behaviors, which may ultimately cause severe damage to the operating environment. In this causal chain, the origin is a failure of security objectives caused by adversaries; however, the subsequent harm caused to the environment falls within the scope of safety objectives. Therefore, ensuring security can oftentimes be a pre-condition to maintaining safety objectives.

2.2 Threat Models

There are also two different threat models that underpin safety and security considerations. Security is predominantly concerned with an adversarial threat model, in which the risks largely result from the deliberate actions of adversaries. Safety, in comparison, has a large focus on non-adversarial scenarios, addressing unintended harms arising from accidents without necessarily assuming the presence of adversaries. Yet, as we elaborated at the end of Section 2.1, security failures due to adversarial attacks may also lead to detrimental behaviors of AI systems that safety objectives aim to prevent. Therefore, adversarial threat models are also becoming an increasingly important part of considerations to achieve safety objectives.

Still, many AI safety works have a biased focus on inherent flaws and unintended faults during the development and operation of AI systems, which may cause accidental harm. For instance, underspecified modeling objectives may result in unintended behaviors in AI systems (D'Amour et al., 2022). Inadequate robustness to distribution shifts could cause failures in less common environments (Hendrycks & Dietterich, 2019). In comparison, the security of AI is predominantly concerned with intentional attacks and exploitations of AI systems by adversaries. For example, attackers can poison training datasets with erroneous data points to cause the trained model to misbehave (Chen et al., 2017; Wan et al., 2023), and adversarial perturbations may be introduced to the model inputs to induce model errors (Szegedy et al., 2014; Goodfellow et al., 2014), and so on.

The two threat models can be contextualized in recent LLM risk studies, illustrating the gaps between the alignment efforts championed by the safety community and the ensuing exposure of adversarial attacks by the security community. Particularly, we note that Supervised Fine-Tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) (Wei et al., 2021; Ouyang et al., 2022; Bai et al., 2022b; Touvron et al., 2023b) have recently been

broadly used to align models with human values, showing promising performance in mitigating many pressing AI safety issues. These are also supplemented by measures such as OpenAI moderation API (OpenAI, 2023a) and Llama Guard (Inan et al., 2023), filtering unsafe model inputs and outputs. However, these safety measures show clear gaps in addressing security threats from adversaries formulated by the security community. For instance, existing alignments are easily circumvented by jailbreaking attacks (Wei et al., 2023; Carlini et al., 2023b; Qi et al., 2023a,c; Zou et al., 2023b), and input and output safety filters remain susceptible to adversarial examples (Zou, 2023).

The ineffectiveness of these safety measures in combatting adversarial attacks is sometimes used to undermine their overall validity. However, a more nuanced understanding can be achieved by evaluating their performance with respect to non-adversarial and adversarial threat models separately. That is – these safety measures do exhibit effectiveness in reducing accidental harm, which is a clear indicator of their efficacy in promoting safety, even though they are not adequately robust against security threats. In this regard, distinguishing between safety and security can be instrumental in defining the appropriate scope and setting accurate expectations concerning the effectiveness of proposed risk mitigation techniques. This position also echoes the separate examination of common corruption and the worst-case adversarial perturbations in visual classification tasks. It has been advocated back in Hendrycks & Dietterich (2019), where they find that a bypassed adversarial defense (i.e., an insecure approach) can still provide substantial common perturbations robustness — an indicator of better safety performance.

2.3 Problem Framing

Now, we also talk about two different types of problem framing in safety and security. When working on a non-adversarial threat model in the safety domain, the risks are generally considered relatively stable over time once identified. Given a sufficiently large dataset, *probabilistic modeling* of accidental safety failures is typically effective, covering the most prevalent fault types and providing reasonably accurate probability estimates — though rare black swans could still occasionally pop out. In contrast, security risks under the adversarial threat model are inherently more dynamic and unpredictable. Adversaries are constantly evolving their attack strategies, and the security of a system is only as strong as its weakest link. Thus, probabilistic estimation of security failures from historical data is less fruitful. A failure route that might be accidentally triggered one in a million times might be 100% exploited by an attacker controlling the route. This nature leads to a preference for a *worst-case modeling* approach for security risks over probabilistic modeling reliant on some static distributions.

This difference is clear in the mathematical formulations in ML settings. Reducing accidents in safety can often be framed as standard risk minimization problems, minimizing a loss function that quantifies the discrepancy between intended and observed outcomes. For instance, current safety alignment approaches are consistently framed as risk minimization. Model vendors collect a large set of red-teaming examples (Ganguli et al., 2022) that enumerate prohibited instructions or questions that users may ask. Safety guardrails are implemented via fine-tuning models to reduce probabilities in responding to these examples, i.e., minimizing the risk on a static distribution estimated by the red-teaming examples. Another instance is CLIP (Radford et al., 2021), which is known to be more robust against common distribution shifts — a good indicator for better safety performance. This could be largely attributed to risk minimization on a web-scale dataset that covers sufficiently many types of common distribution shifts.

In comparison, security objectives are inherently minimax games, where the aim is to minimize the maximum possible loss inflicted by an adversary. Adversarial training (Madry et al., 2017), a technique for defending against adversarial examples, is a classic exemplar. Due to the same rationale, the evaluation of security is also more binary than that of safety. In security studies of AI, as long as a single adaptive attack is found to be able to bypass a defense, the proposed defense is often largely diminished, as we have seen in (Carlini & Wagner, 2016; He et al., 2017; Carlini & Wagner, 2017; Athalye et al., 2018; Tramer et al., 2020; Qi et al., 2022a).

Empirical experience of AI security suggests that the security problems of ML are seldom addressed by safety approaches. Considering the two safety examples from above, safety-aligned LLMs are not robust against jailbreak attacks. Even though CLIP is more robust against common distribution shifts, it neither survives adversarial examples. On the other hand, worst-case performance in some adversarial scenarios can provide a meaningful lower bound for safety performance in average cases. There are safety work benefits using minimax problem framing to do worst-case modeling for improving safety (Hsu et al., 2023).

2.4 Governance and Liability

In conventional contexts, safety and security are often used as different anchor points to group governance systems. Take the aviation industry as an example; safety governance is more concerned with aspects such as the enforcement of stringent safety standards within aircraft facilities and the adherence to procedural guidelines for aircraft operations. These efforts aim to prevent accidents and also minimize harm if accidents do occur. On the

other hand, security is more biased to measures such as passenger and cargo screening, reinforcement of cockpit doors and locks, provision of self-defense weapons, and the presence of air police to deter potential adversaries. In practice, implementing the two governance systems is reliant on two different sets of expertise, resources, and stakeholder engagement. Notably, in the U.S., the Federal Aviation Administration (FAA) and the Transportation Security Administration (TSA) operate as separate entities, with the former's function being biased toward aviation safety and the latter's toward security.

One might argue against drawing such analogies in the context of AI, given the significant differences between AI safety and security and their counterparts in other sectors. But the reality is government agencies are indeed motivated to learn from the experience of safety and security governance in established sectors. For instance, the U.S. National Institute of Standards and Technology (NIST) explicitly states in its AI risk management framework that "*AI safety risk management approaches should take cues from efforts and guidelines in fields such as transportation and healthcare and align with existing sector- or application-specific guidelines or standards*" (NIST, 2023).

In practice, divergent governance practices on safety and security may arise from the different goals of different stakeholders. An example in academia is the differing funding programs targeting the two goals. The recent *Safe Learning-Enabled Systems* program funded by the U.S. National Science Foundation explicitly excludes *"research on securing learning-enabled systems against adversaries"* out of its scope[3]. Meanwhile, the *Guaranteeing AI Robustness against Deception (GARD) program* created by the U.S. Defense Advanced Research Projects Agency (DARPA) has been supporting the development of AI security solutions to defend against adversarial attacks on AI systems[4].

Conflating safety and security may sometimes lead to tricky bias that leaves either security or safety resources on the back burner. An entity biased toward strong adversarial defenses might neglect low-cost mitigations that can be bypassed by a determined adversary (Hendrycks & Dietterich, 2019), even if it reduces the likelihood of accidental failures (Hendrycks & Dietterich, 2019). Similarly, an entity biased toward addressing safety might overlook how strong adversaries would adversarially exploit an AI system to cause large-scale harm (Narayanan & Kapoor, 2023). This imbalance is evident in recent large language model (LLM) technical reports, where AI security considerations are often marginalized by safety priorities (OpenAI, 2023c; Touvron et al., 2023a,b; Gemini Team, 2023). For instance, a keyword search for "safety" in Touvron et al. (2023b) yields 299 matches, while "security" returns only five, all appearing in a general context rather than addressing AI security issues.

Properly disentangling safety and security may have benefits in international negotiation. In past efforts to cooperate in reducing the risks of nuclear weapons, clarifying the differences between nuclear safety and nuclear security issues was an important baseline condition, especially since a single term expresses the meaning of both safety and security in Chinese (*anquan*) and Russian (*bezopasnost*). When the U.S. deliberated about sharing nuclear safety and security techniques with other countries, it was often more feasible to transfer safety technologies because sharing information on nuclear security could expose vulnerabilities in one's own systems (Ding, 2024). Similar distinctions may be helpful for international collaboration in AI risk management as well.

Safety and security may also be perceived and evaluated differently in liability attribution. Safety liability tends to be more clear-cut. In scenarios where a system is used as intended and benignly, the system's producers are more likely to be liable for harm caused by the system's failures. When the operator of the system blatantly disregards established safety standards, attributing liability to the operator's negligence is also straightforward. On the other hand, security liability can be more complex. The benchmarks for what is considered adequate security are constantly evolving alongside the progression of attack and defense techniques. Determining who is responsible for implementing security measures, and to what degree, can often be ambiguous. For example, for AI models, should the model creators or the users be liable for evasion attacks against these models during deployment time? There is a large gray area (Evtimov et al., 2019; Henderson et al., 2023a).

3 Differing Problems Orientation in AI Safety and AI Security

We have discussed the nuances and interplay between AI safety and AI security from a high-level view of some generic principles in Section 2. In this section, we look into some more concrete problems. Specifically, we present an overview of the differing problem focuses within the AI safety and AI security communities by presenting two taxonomies that capture two disjoint sets of problems in academic literature. The first taxonomy (Section 3.1) presents representative problems that have been predominantly investigated by the AI safety community. The second taxonomy (Section 3.2), is characterized by an AI security perspective. The dual taxonomies are intended to provide concrete illustrations of the different mindsets and considerations underlying safety and security. Therefore,

[3] https://www.nsf.gov/pubs/2023/nsf23562/nsf23562.pdf
[4] https://www.darpa.mil/news-events/2019-02-06

they are deliberately structured in non-overlapping manners to highlight the most salient differences. Meanwhile, we also acknowledge that safety and security overlap in many problems and defer the discussion of representative examples to Section 4. We also refer interested readers to an extended version of our taxonomies in Appendices A and B. But note that achieving complete exhaustiveness in these taxonomies is beyond the goal of this paper.

3.1 Representative Safety-Oriented Problems in AI Systems

The central objective of AI safety is to prevent AI systems from causing harm (Section 2.1). Surrounding this objective, numerous pivotal problems in AI safety have been primarily examined in non-adversarial settings (Section 2.2). This subsection elaborates on this constrained set of problems. We group these problems along two major axes: **1) safe usability** (Section 3.1.1), referring to the safety-critical qualities that are necessary for AI systems to be safe to use; and **2) safe operation** (Section 3.1.2), denoting the generic safety principles that should be adhered to during the operation of AI systems. Meanwhile, a few other AI safety problems that involve considerations of adversarial threat models are deferred to a joint discussion with AI security in Section 4.

3.1.1 Safe Usability

We use *safe usability* to categorize the qualities an AI system should have to make them safe to use under some intended conditions. Many representative AI safety problems can be reduced to such qualities. We list a few below.

Robustness to Distributional Shift. Training data of an AI model does not perfectly represent the distribution of the real-world environment where the model will be applied. There could be a long-tail distribution of unusual events in the real world (Hendrycks & Dietterich, 2019; Taleb, 2020) that are not seen during training. The distribution could also drift over time because the world is changing by itself (Bayram et al., 2022). AI systems should be prepared for such distribution shifts and avoid causing catastrophic harm in out-of-distribution edge cases.

Calibration. AI systems should have the capability to calibrate their confidence with their answers. In classification, this means predicting probability estimates representative of the true correctness likelihood (Guo et al., 2017). This principle should also extend to recent LLMs, ensuring that the model remains conservative stances or abstains from providing an answer when it is uncertain. Calibration ensures users are well-informed about the reliability of the outcomes produced by the AI systems, prompting human discretion when the results are less dependable.

Honesty and Truthfulness. AI systems should be honest when they know what the real answers are (Li et al., 2023b). They should also be truthful, hold the right knowledge of factuality, and avoid hallucination (Lin et al., 2021).

Alignment. AI systems should be aligned with human values. For example, they should be able to reject inappropriate instructions, such as requests to generate disinformation, child abuse content, or partake in other unethical or unlawful activities (Bai et al., 2022b,c). They should also demonstrate respect towards minority groups by maintaining fairness and impartiality and avoiding offensive, discriminatory content or hate speech (Gehman et al., 2020). More generally, alignment means that the behaviors of AI systems are aligned with the real (or preferred) objectives that humans intend to achieve (Gabriel, 2020). Counter-examples are the situations where an AI system does exactly fulfill a given objective but in a completely unintended way. This can include reward-hacking (Pan et al., 2022); for example, a robot achieves an environment free of messes by disabling its vision so that it won't find any messes. Another example is the side effects in which AI systems achieve the objective but simultaneously create negative consequences while carrying out the given objective (Amodei et al., 2016).

Interpretability and Explainability. Interpretability of AI systems focuses on understanding the inner workings of the models, while explainability intends to explain the final decisions made by AI systems (AWS, 2021). They are closely relevant and are thus sometimes jointly denoted under the umbrella term of AI transparency[5] (Zou et al., 2023a). Both of the two properties are now deemed increasingly safety-critical. Highly interpretable inner workings enable the timely discovery of anomaly behaviors and functionalities of AI systems and thus can facilitate the prevention of hazards (Hendrycks et al., 2021b). More explainable AI decisions enable humans to understand, appropriately trust, and more effectively manage AI systems (Xu et al., 2019), increasing accountability.

3.1.2 Safe Operation

Just like car drivers need to obey traffic rules to drive safely, AI safety is also about safely operating AI systems.

Ensuring Correct Contexts of Application. A model should not be applied to a task that it is not qualified to fulfill. For example, a conversational agent like ChatGPT does not have the professional qualification to provide medical,

[5]The term "transparency" nowadays is also used to denote the transparency of the development process of AI models (Bommasani et al., 2023) beyond just the transparency of the working mechanism of AI models.

mental, or legal advice. It was neither developed for nor rigorously tested for such professional services (OpenAI, 2024). Applying a model out of the intended contexts may get inaccurate or undesired results, which in turn could harm the users. AI service providers are responsible for documenting the model's purpose and clarifying the correct contexts of application and the model's limitations. They are also responsible for making users well-informed, as many ordinary users of AI services may not know these issues.

Avoiding Overreliance. When AI systems become increasingly more capable, users may excessively trust and depend on them without exerting necessary caution and oversight (OpenAI, 2023c). Lawyers cite fake cases generated by ChatGPT is a representative example (Reed, 2023). Users should be educated about the limitations of AI systems, and policies should be in place to restrict people from overly relying on AI systems in some critical contexts.

Controllable Autonomy. There is also a concern that increasingly powerful AI systems may get out of our control if they are operated in an overly autonomous manner. For instance, a learning agent may learn to prevent operators from shutting it down, as this interruption hinders its ability to accomplish initially assigned tasks (Orseau & Armstrong, 2016). Another hypothetical situation is power-seeking behavior (Hendrycks et al., 2023; Carlsmith, 2022), in which the AI system persistently seeks power and control over the environment to achieve its assigned objectives, despite posing a direct threat to societal stability and existence. Although these risks are inherently speculative and futuristic in nature, they underscore the importance of operators carefully evaluating their ability to maintain control over an AI agent when granting it increasing autonomy.

3.2 Representative Security-Oriented Problems in AI Systems

AI security works on adversarial threat models (Section 2.2) and focuses on protecting security-critical properties of AI systems from being compromised (Section 2.1). We now discuss representative AI security properties and common attacks that could threaten them. Particularly, we elaborate on the CIA Triad, i.e., confidentiality, integrity, and availability, that form the basis of information security. This provides a breakdown illustration of established AI security considerations, which share clearly different mindsets compared with the safety-oriented viewpoints.[6]

3.2.1 Confidentiality

Confidentiality denotes the assurance that information and data with access restrictions are not made accessible to unauthorized entities or processes. Many aspects of AI systems can involve confidentiality.

A typical example is *training data* that contains proprietary or sensitive records and thus should not be disclosed. However, neural network models often excessively memorize their training data (Carlini et al., 2019), rendering the data subject to privacy attacks. Adversaries can apply membership inference attack (MIAs) (Shokri et al., 2016) to infer whether a specific data record was used in training a model by only observing the model's prediction on this data record. Another threat is data extraction attacks (Carlini et al., 2021) that can directly recover large chunks of training data verbatim from trained models, even including the latest production-level chatbots (Nasr et al., 2023a).

System prompts of many LLMs are now considered secrets as well, due to varying reasons. For example, a good system prompt often needs a lot of prompt engineering and could be a type of intellectual property. These prompts may also just contain private in-context training data (Tang et al., 2024). Confidentiality of system prompts is threatened by prompt extraction attacks (Zhang & Ippolito, 2023). Recently, system prompts of custom models in GPTs ecosystem were broadly leaked by such attacks (Leaked-GPTs, 2024).

Users' interactions with AI systems constitute another type of confidential information. Logs with chatbot-style AI systems contain a lot of sensitive information in both personal and enterprise usage. Adversarial threats in this context are on the rise. Sometimes, third-party AI service providers themselves may act as adversaries, as they can access all server logs. Actors in the wild can also represent a threat: recently, Rehberger (2024) showed the indirect prompt injection can cause Amazon's Q model (for business usage) to generate malicious URLs during the chat, which, once clicked, will send chat history to attackers' servers.

[6] We didn't discuss operational-related AI security issues like we did for AI safety in Section 3.1.2, as they are less examined in current AI security contexts. Nonetheless, it is worth mentioning that, in conventional information security, awareness of security during operation plays a significant role in overall security considerations. For example, key management can be as important as encryption mechanisms because poor practices (e.g., using weak passwords or sharing passwords) can equally render the confidentiality property more susceptible to adversarial attacks. Another typical example is patch management. Users should regularly update and patch operating systems, applications, and software. Failing to do so can leave systems vulnerable to many known exploits and malware. As AI systems are being deployed in more security-critical applications, we expect that operational-related issues will also be an important future direction for AI security considerations.

3.2.2 Integrity

Security is also about integrity, a concept including both *system integrity* and *data integrity* (Nieles et al., 2017).

System integrity is the quality that a system has when it performs its intended function in an unimpaired manner, free from unauthorized manipulation. The system integrity of AI systems is challenged by *evasion attacks*, which denote the scenario where an adversary manipulates the input sample to an ML model to cause erroneous model inferences (Biggio et al., 2013). *Adversarial examples* for classification models represent a typical evasion attack (Szegedy et al., 2014; Goodfellow et al., 2014), where attackers apply small perturbations to normal samples, causing the model to generate wrong classifications. For the latest advanced AI systems with more complicated functions beyond mere classification, both the range of system integrity requirements and the scope for potential adversarial threats expand. *Jailbreak attacks* on aligned LLMs are notable examples where attackers can manipulate inputs to break the integrity of the LLMs' safety guardrails (Qi et al., 2023a; Zou et al., 2023b), such that they can no longer prevent harmful behaviors of models. Another instance is *prompt injection attacks* (Liu et al., 2023b; Abdelnabi et al., 2023; Liu et al., 2023a) on LLMs integrated systems. Attackers manipulate model inputs to induce LLMs to generate malicious outputs that can harm the broader systems that consume these outputs.

System integrity can also be compromised by attackers who poison an AI model's training dataset (Goldblum et al., 2022), e.g., manipulating a small portion of training data points. This can result in the trained model producing errors on specific test examples (Shafahi et al., 2018) or embedded with prohibited backdoor behaviors (Chen et al., 2017; Yan et al., 2023; Xu et al., 2023a; Hubinger et al., 2024).

Data integrity is the property that data and information are not altered in an unauthorized manner. For AI systems, this involves the protection of training data, model weights, and other artifacts like codes, dependent libraries, configuration files, and so on, from unauthorized modification or destruction. When AI systems are used to oversee other systems (e.g., external databases), protecting the integrity of data of those connected systems can also be tightly relevant (Pedro et al., 2023).

3.2.3 Availability

Availability is the property that a system and its resources are accessible and usable on demand by an authorized entity. In AI systems, this can be further divided into *training-stage availability* and *inference-stage availability*.

Training-stage availability means that a training procedure can produce a valid and usable model, free from adversarial disturbance that would make the trained model invalid. A common threat to training-stage availability is indiscriminate data poisoning attacks (Nelson et al., 2008; Biggio et al., 2012; Muñoz-González et al., 2017; Lu et al., 2023). In such attacks, adversaries moderately manipulate the training dataset, resulting in trained models with substantially degraded performance, essentially rendering them unusable.

Inference-stage availability requires that AI systems provide services on demand in a timely manner. It can be threatened by resource depletion attacks, where adversaries launch large volumes of dummy service requests to an AI system such that it can not serve normal users. This can be further combined with techniques like sponge examples (Shumailov et al., 2021) that are model inputs designed to maximize energy consumption and latency. Resource depletion attacks can be particularly threatening for large foundation models that need increasingly more computation resources. Besides, prompt injection attacks (Abdelnabi et al., 2023) have recently also been used to compromise the inference-stage availability of LLMs. Basically, if some adversarial prompt can be injected, the models can be instructed to do dummy tasks or just ignore real tasks.

4 Bridging AI Safety and AI Security in Frontier AI Risk Management

So far, we have elucidated how safety and security have conventionally shaped differing considerations and problem focuses. But looking ahead, we highlight that many practical AI risk management challenges also necessitate bridging and unifying safety and security perspectives for effective resolution.

Case Study 1: Safety and Security in Combating Malicious Use. Combating the malicious use of advanced AI models is a prominent scenario in which both safety and security matter. In the first place, the malicious use of AI directly opposes the objective of AI safety, which aims to prevent AI systems from causing harm (Section 2.1). The safety community has thus proposed alignment approaches (Ouyang et al., 2022; Bai et al., 2022b; Leike & Sutskever, 2023), which involve the fine-tuning of models to ensure their alignment with human values (Section 3.1.1). These aligned AI systems gain the ability to reject harmful instructions from malicious actors who seek to misuse them. The alignment approach is now widely embraced as a core mechanism for implementing safety guardrails in almost all major commercial LLMs. However, the effectiveness of this method relies on relatively benign scenarios where

adversaries do not actively retaliate. In a realistic adversarial threat model (Section 2.2) from a security perspective, malicious actors are unlikely to send merely plain harmful instructions to an aligned model and naively wait for rejection. Instead, determined adversaries would consistently strive to circumvent the safety guardrail provided by the alignment approach, compelling AI models to facilitate malicious use despite the alignment (Wei et al., 2023; Zou et al., 2023b; Qi et al., 2023c; Narayanan & Kapoor, 2023). Thus, combating malicious use effectively extends beyond simply aligning AI with human values and implementing safety guardrails; it also necessitates the protection of integrity (Section 3.2.2) of the safety guardrail itself against adversarial attacks, which corresponds to a separate line of security efforts in designing adversarial defenses (Jain et al., 2023; Robey et al., 2023; Zhou et al., 2024). This also has immediate implications for the practice of red-teaming. Early red-teaming on LLMs (Ganguli et al., 2022) seldom considers adversarial attacks that have been commonly considered by AI security. We suggest practitioners be more explicit on what the threat model is. If the goal is to combat malicious use by determined adversaries, security perspectives should definitely be more seriously considered in red-teaming (Mazeika et al., 2024).

Case Study 2: Safety and Security in The Transparency of AI Generated Content (AIGC). *From a safety perspective*, as AI systems become more prevalent and sophisticated, it's also crucial that they are transparently identified as non-human entities. This principle of transparency ensures that individuals are aware they're interacting with an AI, fostering informed consent and engagement. Consistent with recent US executive orders, synthetic data such as machine-generated text or images should be clearly marked, enhancing user awareness and accountability (Biden, 2023). For instance, ChatGPT implemented this by explicitly stating their AI nature in communications, aiding users, especially non-technical ones, in proper identification. This aspect of identifiability is essential not only for distinguishing between human-generated and machine-generated content but also for mitigating the spread of spam and misinformation. While social media platforms are particularly vulnerable to the misrepresentation of machine-generated content, strategies like overt labeling and sophisticated techniques, including imperceptible watermarking, can significantly enhance transparency and mitigate harm (Christ et al., 2023). *Yet, from a strict security game perspective*, the strict detectability of AIGC, in general, is an exceedingly hard problem. For example, in the context of watermarking, attacks such as oracle attacks (Zhang et al., 2023) or generative attacks (Kirchenbauer et al., 2023) can be used to adversarially invalidate the watermark and thus rendering AIGC indetectable. This is especially apparent in domains such as text where, theoretically, for every watermarked document, a semantically equivalent, unwatermarked document exists, which an attacker with sufficient information about the system can exploit. This discrepancy between safety and security is a prominent source of confusion in both academic literature and policy concerning AIGC, where, for example, policymakers create legislation aimed at AIGC well-suited for safety goals, but expect this to lead to simple resolutions for hard questions in security as well.

The above two examples highlight the importance of considering both safety and security perspectives in practical AI risk management. In general, **security can be a critical precondition for safety**, as the safety measures themselves risk being adversarially invalidated by determined adversaries that need security in place to defend against. If the consequences of failures are bound to be catastrophic, robust security must be established to ensure that these harms remain bounded even in the worst cases (Yampolskiy & Spellchecker, 2016).

Unfortunately, **safety and security objectives can also clash**. We have discussed in Section 2.3 that the evaluation of security is more binary than that of safety. Security is biased to lower-bound worst cases by its nature, while good performance in average cases is a baseline requirement for safety. However, sometimes worst-case and average-case performances may not be simultaneously optimizable. In the context of watermarking AIGC, perfect security for AIGC detectability might be theoretically unattainable (Sadasivan et al., 2023), but the watermark can still be beneficial in average cases without adversaries, in favor of safety objectives. More often than not, numerous adversarial threat models encountered by ML models present worse explicit trade-offs between safety and security. Adversarial training (Madry et al., 2017) and randomized smoothing (Cohen et al., 2019) have been shown to be effective in enhancing worst-case adversarial robustness (improving security) while sacrificing benign accuracy in average cases (Tsipras et al., 2018), which can render the models less safe to use. Similarly, Moayeri et al. (2022) demonstrates the trade-off between adversarial robustness and natural distribution robustness. These phenomena may continue to serve as fundamental technical challenges in balancing safety and security practices.

Stakeholders should be fully aware of these dependencies as well as the technical tension between the practices of AI safety and AI security. Developing a comprehensive understanding of these elements is imperative for bridging the two sets of efforts and fostering a balanced, holistic risk management approach. Specifically, as government agencies are starting to initiate their efforts in managing AI risks, as reflected in the recent establishment of UK AI Safety Institute[7] by the UK government and the U.S. AI Safety Institute (USAISI)[8] led by NIST, these understandings are particularly urgent for the ensuing operational items.

[7] https://www.gov.uk/government/publications/ai-safety-institute-overview

[8] https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

5 Conclusion

In this paper, we advocate that stakeholders in AI risk management should unambiguously understand and take into account both safety and security perspectives in their practices. We note that this vision is often obscured by the poor and inconsistent definitions of the concepts of safety and security themselves. To address this conceptual challenge, we propose a reference framework to facilitate a common understanding of the differences and interplay between safety and security in their objectives of protection, threat models, and problem framing, as well as promoting a more nuanced understanding of their implications in governance and liability structures. Based on the nuanced understanding of the safety and security perspectives we have established, we analyze the importance of bridging AI safety and AI security perspectives in addressing frontier AI risk management problems while also discussing how the two can sometimes technically clash. By clarifying the two critical concepts, this paper aims to contribute to the development of more holistic evaluation, investments, and incentives, as well as overall risk management.

More broadly, we also acknowledge the relevance of other critical disciplines in AI risk management that are not comprehensively covered within this paper. As our focus primarily lies in safety and security perspectives, these topics are less examined in our discussions — but we supplement an overview on a few of them in Appendix C.

Broader Impacts

Throughout this work, we emphasize a reference framework that would provide sufficient emphasis on both security and safety considerations. We believe this is an important reframing of considerations that would ensure consistent coverage in evaluation frameworks, consistent investment, and overall better risk management by taking on a holistic perspective spanning multiple disciplines. We believe this would improve on the status quo, particularly as policymakers put in place structures to evaluate and regulate systems, leading to a broader positive impact. Even if others disagree, we believe this discussion is worth engaging in more broadly, bringing in additional stakeholders from other perspectives and other disciplines.

Acknowledgements

We thank Dan Hendrycks , Sadhika Malladi, Ashwinee Panda, Mengzhou Xia, Chiyuan Zhang, and Andy Zou for providing helpful feedback on this paper. We thank the great support from Princeton LLM Alignment reading group, Princeton Language and Intelligence (PLI), and Princeton Center for Information Technology Policy (CITP). Prateek Mittal acknowledges the support by NSF grants CNS-2131938 and Princeton SEAS Innovation Grant. Bo Li acknowledges the support from the NSF grant No. 2046726, DARPA GARD, and NASA grant no. 80NSSC20M0229, Alfred P. Sloan Fellowship, the Amazon research award, and the eBay research grant. Ruoxi Jia and the ReDS lab acknowledge support through grants from the Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, the NSF grant No. IIS-2312794, NSF IIS-2313130, NSF OAC-2239622, and the CCI SWVA Research Engagement Award. Xiangyu Qi is supported by the Princeton Gordon Y. S. Wu Fellowship. Yangsibo Huang is supported by the Wallace Memorial Fellowship. Edoardo Debenedetti is supported by armasuisse Science and Technology.

References

Abdelnabi, S., Greshake, K., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023.

Abid, A., Farooqi, M., and Zou, J. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Administration, U. G. S. Rules and Policies - Protecting PII - Privacy Act. https://www.gsa.gov/reference/gsa-privacy-program/rules-and-policies-protecting-pii-privacy-act, 2019.

Aghakhani, H., Meng, D., Wang, Y.-X., Kruegel, C., and Vigna, G. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 159–178. IEEE, 2021.

Aïvodji, U., Bolot, A., and Gambs, S. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

Altman, O. Planning for agi and beyond. https://openai.com/blog/planning-for-agi-and-beyond, 2024.

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., and Consortium, P. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9, 2020.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anderson, R. Security Engineering Lecture 13: Safety and Security. https://www.youtube.com/watch?v=uZkQtnHKcJ4, 2022.

Anthropic. Introducing Claude. https://www.anthropic.com/index/introducing-claude, 2023.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., and Felici, G. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. J. Secur. Netw.*, 10(3):137–150, sep 2015. ISSN 1747-8405. doi:10.1504/IJSN.2015.071829. URL https://doi.org/10.1504/IJSN.2015.071829.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

AWS. Interpretability versus explainability. https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html, 2021.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2938–2948. PMLR, 26–28 Aug 2020. URL http://proceedings.mlr.press/v108/bagdasaryan20a.html.

Bai, J., Wu, B., Zhang, Y., Li, Y., Li, Z., and Xia, S.-T. Targeted attack against deep neural networks via flipping limited weight bits. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iKQAk8a2kM0.

Bai, J., Gao, K., Gong, D., Xia, S.-T., Li, Z., and Liu, W. Hardly perceptible trojan attack against neural networks with bit flips. In *European Conference on Computer Vision*, pp. 104–121. Springer, 2022a.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022b.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022c.

Balunovic, M., Dimitrov, D., Jovanović, N., and Vechev, M. Lamp: Extracting text from gradients with language model priors. *Advances in Neural Information Processing Systems*, 35:7641–7654, 2022.

Barbalau, A., Cosma, A., Ionescu, R. T., and Popescu, M. Black-box ripper: Copying black-box models using generative evolutionary algorithms. *Advances in Neural Information Processing Systems*, 33:20120–20129, 2020.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

Barreno, M., Nelson, B., Joseph, A. D., and Tygar, J. D. The security of machine learning. *Machine Learning*, 81: 121–148, 2010.

Bayardo, R. J. and Agrawal, R. Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pp. 217–228. IEEE, 2005.

Bayram, F., Ahmed, B. S., and Kassler, A. From concept drift to model degradation: An overview on performance-aware drift detectors, 2022.

Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, 2023.

Biden, J. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, 2023.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pp. 387–402. Springer, 2013.

Binder, A., Bockmayr, M., Hägele, M., Wienert, S., Heim, D., Hellweg, K., Ishii, M., Stenzinger, A., Hocke, A., Denkert, C., et al. Morphological and molecular breast cancer profiling through explainable machine learning. *Nature Machine Intelligence*, 3(4):355–366, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023.

Bostrom, N. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford, 2014.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lovitt, L., Elhage, N., Schiefer, N., Joseph, N., Mercado, N., DasSarma, N., Larson, R., McCandlish, S., Kundu, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., and Kaplan, J. Measuring progress on scalable oversight for large language models, 2022.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Breier, J., Hou, X., Jap, D., Ma, L., Bhasin, S., and Liu, Y. Practical fault attack on deep neural networks. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2204–2206, 2018.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

Burns, A., McDermid, J., and Dobson, J. On the meaning of safety and security. *The computer journal*, 35(1):3–15, 1992.

Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q. A., Fu, K., and Mao, Z. M. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 2267–2281, 2019.

Carlini, N. and Terzis, A. Poisoning and backdooring contrastive learning. In *ICLR*, 2022.

Carlini, N. and Wagner, D. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 3–14, 2017.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Carlini, N., Jagielski, M., and Mironov, I. Cryptanalytic extraction of neural network models. In *Annual International Cryptology Conference*, pp. 189–218. Springer, 2020.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022a.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022b.

Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., and Tramer, F. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022c.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023a.

Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Awadalla, A., Koh, P. W., Ippolito, D., Lee, K., Tramer, F., et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023b.

Carlsmith, J. Is power-seeking ai an existential risk?, 2022.

Carranza, A., Pai, D., Schaeffer, R., Tandon, A., and Koyejo, S. Deceptive alignment monitoring. *arXiv preprint arXiv:2307.10569*, 2023.

Chabanne, H., Danger, J.-L., Guiga, L., and Kühne, U. Side channel attacks for architecture extraction of neural networks. *CAAI Transactions on Intelligence Technology*, 6(1):3–16, 2021.

Chang, H., Nguyen, T. D., Murakonda, S. K., Kazemi, E., and Shokri, R. On adversarial bias and the robustness of fair machine learning. *arXiv preprint arXiv:2006.08669*, 2020.

Chen, L., Zaharia, M., and Zou, J. How is chatgpt's behavior changing over time?, 2023a.

Chen, P.-Y. and Das, P. AI maintenance: A robustness perspective. *Computer*, 56(2):48–56, 2023.

Chen, P.-Y. and Hsieh, C.-J. *Adversarial Robustness for Machine Learning*. Academic Press, 2022.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. In *arXiv:1712.05526*, 2017.

Chen, X., Tang, S., Zhu, R., Yan, S., Jin, L., Wang, Z., Su, L., Wang, X., and Tang, H. The janus interface: How fine-tuning in large language models amplifies the privacy risks. *arXiv preprint arXiv:2310.15469*, 2023b.

Chen, Y., Shen, C., Shen, Y., Wang, C., and Zhang, Y. Amplifying membership exposure via data poisoning. *Advances in Neural Information Processing Systems*, 35:29830–29844, 2022.

Christ, M., Gunn, S., and Zamir, O. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Cole, S. ChatGPT Users Report Seeing Other People's Conversation Histories . https://www.vice.com/en/article/5d9zd5/chatgpt-users-report-being-able-to-see-random-peoples-chat-histories, 2023.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Huang, Y., Jagielski, M., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y.-X., Xiong, L., Yekhanin, S., Yu, D., Zhang, H., and Zhang, W. Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment. *Harvard Data Science Review*, jan 16 2024. https://hdsr.mitpress.mit.edu/pub/sl9we8gh.

Dalvi, N., Domingos, P., Mausam, Sanghai, S., and Verma, D. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108, 2004.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, 23(1):10237–10297, 2022.

Dastin, J. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pp. 296–299. Auerbach Publications, 2022.

Deng, J. and Ma, J. Computational copyright: Towards a royalty model for ai music generation platforms. *arXiv preprint arXiv:2312.06646*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Deng, Y., Zhang, W., Pan, S. J., and Bing, L. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

Ding, J. Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies . https://jeffreyjding.github.io/documents/Keep%20Your%20Enemies%20Safer%20Nov%202022.pdf, 2024.

Drapkin, A. Does ChatGPT Save My Data? OpenAI's Privacy Policy Explained. https://tech.co/news/does-chatgpt-save-my-data, 2023.

Duan, J., Kong, F., Wang, S., Shi, X., and Xu, K. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

Dubiński, J., Kowalczuk, A., Pawlak, S., Rokita, P., Trzciński, T., and Morawiecki, P. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4860–4869, 2024.

Duddu, V., Samanta, D., Rao, D. V., and Balas, V. E. Stealing neural networks via timing side channels. *arXiv preprint arXiv:1812.11720*, 2018.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Elsayed, G. F., Goodfellow, I., and Sohl-Dickstein, J. Adversarial reprogramming of neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Syx_Ss05tm.

European Commission. Eu artificial intelligence act. https://artificialintelligenceact.eu/the-act/, 2023.

European Parliament and Council of the European Union. Health Insurance Portability and Accountability Act of 1996. https://www.govinfo.gov/app/details/PLAW-104publ191, 1997.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj, 2016.

Evtimov, I., O'Hair, D., Fernandes, E., Calo, R., and Kohno, T. Is tricking a robot hacking? *Berkeley Technology Law Journal*, 34(3):891–918, 2019.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. Technical report, CEPR Discussion Papers, 2017.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., and Borisov, N. 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 '18, pp. 619–633, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356930. doi:10.1145/3243734.3243834. URL https://doi.org/10.1145/3243734.3243834.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ghazi, B., Huang, Y., Kamath, P., Kumar, R., Manurangsi, P., Sinha, A., and Zhang, C. Sparsity-preserving differentially private training of large embedding models. *arXiv preprint arXiv:2311.08357*, 2023.

Ghosal, S. S., Chakraborty, S., Geiping, J., Huang, F., Manocha, D., and Bedi, A. A survey on the possibilities & impossibilities of AI-generated text detection. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=AXtFeYjboj. Survey Certification.

Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Madry, A., Li, B., and Goldstein, T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Goldreich, O. Secure multi-party computation. *Manuscript. Preliminary version*, 78(110):1–108, 1998.

Gong, X., Wang, Q., Chen, Y., Yang, W., and Jiang, X. Model extraction attacks and defenses on cloud-based machine learning models. *IEEE Communications Magazine*, 58(12):83–89, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

GOV.UK. Ai safety summit 2023. https://www.gov.uk/government/topical-events/ai-safety-summit-2023, 2023.

Greenberg, J. Studying organizational justice cross-culturally: fundamental challenges. *International Journal of Conflict Management*, 12(4):365–375, 2001.

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., and Weller, A. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*, pp. 903–912, 2018.

Grzybowski, A., Jin, K., and Wu, H. Challenges of artificial intelligence in medicine and dermatology. *Clinics in Dermatology*, 2024.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Gupta, S., Huang, Y., Zhong, Z., Gao, T., Li, K., and Chen, D. Recovering private text in federated learning of language models. *Advances in Neural Information Processing Systems*, 35:8130–8143, 2022.

He, W., Wei, J., Chen, X., Carlini, N., and Song, D. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th USENIX workshop on offensive technologies (WOOT 17)*, 2017.

He, Y., Meng, G., Chen, K., Hu, X., and He, J. {DRMI}: A dataset reduction technology based on mutual information for black-box attacks. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1901–1918, 2021.

Henderson, P., Hashimoto, T., and Lemley, M. Where's the liability in harmful ai speech? *arXiv preprint arXiv:2308.04635*, 2023a.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023b.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021b.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic ai risks, 2023.

Hoepner, A. G., McMillan, D., Vivian, A., and Wese Simen, C. Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective. *The European Journal of Finance*, 27(1-2):1–7, 2021.

Hong, D. K., Kloosterman, J., Jin, Y., Cao, Y., Chen, Q. A., Mahlke, S., and Mao, Z. M. Avguardian: Detecting and mitigating publish-subscribe overprivilege for autonomous vehicle systems. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 445–459. IEEE, 2020.

Hsu, K.-C., Nguyen, D. P., and Fisac, J. F. Isaacs: Iterative soft adversarial actor-critic for safety. In *Learning for Dynamics and Control Conference*, pp. 90–103. PMLR, 2023.

Huang, J., Shao, H., and Chang, K. C.-C. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022a.

Huang, K., Altosaar, J., and Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., and Tygar, J. D. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pp. 43–58, 2011.

Huang, W. R., Geiping, J., Fowl, L., Taylor, G., and Goldstein, T. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33:12080–12091, 2020.

Huang, Y., Huang, C.-Y., Li, X., and Li, K. A dataset auditing method for collaboratively trained machine learning models. *IEEE Transactions on Medical Imaging*, 2022b.

Huang, Y., Gupta, S., Xia, M., Li, K., and Chen, D. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023a.

Huang, Y., Gupta, S., Zhong, Z., Li, K., and Chen, D. Privacy implications of retrieval-based language models. *arXiv preprint arXiv:2305.14888*, 2023b.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., Sachan, K., Sellitto, M., Sharma, M., DasSarma, N., Grosse, R., Kravec, S., Bai, Y., Witten, Z., Favaro, M., Brauner, J., Karnofsky, H., Christiano, P., Bowman, S. R., Graham, L., Kaplan, J., Mindermann, S., Greenblatt, R., Shlegeris, B., Schiefer, N., and Perez, E. Sleeper agents: Training deceptive llms that persist through safety training, 2024.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Investor, T. P. Trick ChatGPT to say its SECRET PROMPT . https://medium.com/@pareto_investor/trick-chatgpt-to-say-its-secret-prompt-973990a27b11, 2023.

Iqbal, U., Kohno, T., and Roesner, F. Llm platform security: Applying a systematic evaluation framework to openai's chatgpt plugins. *arXiv preprint arXiv:2309.10254*, 2023.

Ito, S., Harada, R., and Kikuchi, H. Risk of re-identification from payment card histories in multiple domains. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pp. 934–941. IEEE, 2018.

Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., and Papernot, N. High accuracy and high fidelity extraction of neural networks. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1345–1362, 2020a.

Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020b.

Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P.-y., Goldblum, M., Saha, A., Geiping, J., and Goldstein, T. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Jeong, C., Jang, S., Park, E., and Choi, S. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124:1907–1922, 2020.

jutogashi. Gpt3.5 just adding a random dude's photo in the reply. https://www.reddit.com/r/ChatGPT/comments/17krsg2/gpt35_just_adding_a_random_dudes_photo_in_the/, 2023.

Juuti, M., Szyller, S., Marchal, S., and Asokan, N. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 512–527. IEEE, 2019.

Kahn, D. *The Codebreakers: The comprehensive history of secret communication from ancient times to the internet*. Simon and Schuster, 1996.

Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., and Oh, S. J. Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881*, 2023.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Kohli, P. and Chadha, A. *Enabling Pedestrian Safety Using Computer Vision Techniques: A Case Study of the 2018 Uber Inc. Self-driving Car Crash*, pp. 261–279. Springer International Publishing, February 2019. ISBN 9783030123888. doi:10.1007/978-3-030-12388-8_19. URL http://dx.doi.org/10.1007/978-3-030-12388-8_19.

Korbak, T., Elsahar, H., Kruszewski, G., and Dymetman, M. Controlling conditional language models without catastrophic forgetting, 2022.

Korbak, T., Shi, K., Chen, A., Bhalerao, R. V., Buckley, C., Phang, J., Bowman, S. R., and Perez, E. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pp. 17506–17533. PMLR, 2023.

Kreps, S., McCain, R. M., and Brundage, M. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117, 2022.

Krishna, K., Tomar, G. S., Parikh, A. P., Papernot, N., and Iyyer, M. Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366*, 2019.

Krishna, K., Song, Y., Karpinska, M., Wieting, J. F., and Iyyer, M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=WbFhFvjjKj.

Kuditipudi, R., Thickstun, J., Hashimoto, T., and Liang, P. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Kunar, M. A. and Watson, D. G. Framing the fallibility of computer-aided detection aids cancer detection. *Cognitive Research: Principles and Implications*, 8(1):30, 2023.

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

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. Towards a science of human-ai decision making: A survey of empirical studies, 2021.

Lazar, S. and Nelson, A. Ai safety on whose terms?, 2023.

Leaked-GPTs. Gpts prompts leaked list. https://github.com/friuns2/Leaked-GPTs, 2024.

Leike, J. and Sutskever, I. Introducing Superalignment. https://openai.com/blog/introducing-superalignment, 2023.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction, 2018.

Lermen, S., Rogers-Smith, C., and Ladish, J. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b. *arXiv preprint arXiv:2310.20624*, 2023.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Li, C., Pang, R., Xi, Z., Du, T., Ji, S., Yao, Y., and Wang, T. An embarrassingly simple backdoor attack on self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4367–4378, 2023a.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023b.

Li, S., Wang, X., Xue, M., Zhu, H., Zhang, Z., Gao, Y., Wu, W., and Shen, X. S. Yes, one-bit-flip matters! universal dnn model inference depletion with runtime code fault injection. *Proceedings of the 33th USENIX Security Symposium*, 2024.

Li, Y., Jiang, Y., Li, Z., and Xia, S.-T. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Lin, Y., Tan, L., Lin, H., Zheng, Z., Pi, R., Zhang, J., Diao, S., Wang, H., Zhao, H., Yao, Y., et al. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*, 2023a.

Lin, Z., Xu, K., Fang, C., Zheng, H., Ahmed Jaheezuddin, A., and Shi, J. Quda: query-limited data-free model extraction. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, pp. 913–924, 2023b.

Liu, T., Deng, Z., Meng, G., Li, Y., and Chen, K. Demystifying rce vulnerabilities in llm-integrated apps. *arXiv preprint arXiv:2309.02926*, 2023a.

Liu, Y., Wei, L., Luo, B., and Xu, Q. Fault injection attack on deep neural network. In *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 131–138. IEEE, 2017.

Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., and Zhang, X. Trojaning attack on neural networks. In *NDSS*, 2018.

Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., Wang, H., Zheng, Y., and Liu, Y. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023b.

Lowd, D. and Meek, C. Good word attacks on statistical spam filters. In *CEAS*, volume 2005, 2005.

Lu, Y., Kamath, G., and Yu, Y. Exploring the limits of model-targeted indiscriminate data poisoning attacks. In *International Conference on Machine Learning*, pp. 22856–22879. PMLR, 2023.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083*, 2017.

Mahloujifar, S., Inan, H. A., Chase, M., Ghosh, E., and Hasegawa, M. Membership inference on word embedding and beyond. *arXiv preprint arXiv:2106.11384*, 2021.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., and Halgamuge, M. N. From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*, 2023.

Mehrabi, N., Naveed, M., Morstatter, F., and Galstyan, A. Exacerbating algorithmic bias through fairness attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8930–8938, 2021.

Meta, 2023. URL https://ai.meta.com/llama/use-policy/.

Milli, S., Hadfield-Menell, D., Dragan, A., and Russell, S. Should robots be obedient? *arXiv preprint arXiv:1705.09990*, 2017.

Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. Silo language models: Isolating legal risk in a nonparametric datastore. *arXiv preprint arXiv:2308.04430*, 2023.

Moayeri, M., Banihashem, K., and Feizi, S. Explicit tradeoffs between adversarial and natural distributional robustness. *Advances in Neural Information Processing Systems*, 35:38761–38774, 2022.

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Michi, A., et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., and Roli, F. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 27–38, 2017.

Narayanan, A. and Kapoor, S. Model alignment protects against accidental harms, not intentional ones. https://www.aisnakeoil.com/p/model-alignment-protects-against, 2023.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models, 2023a.

Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., and Terzis, A. Tight auditing of differentially private machine learning. *arXiv preprint arXiv:2302.07956*, 2023b.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I., Saini, U., Sutton, C., Tygar, J. D., and Xia, K. Exploiting machine learning to subvert your spam filter. *LEET*, 8(1-9):16–17, 2008.

Nieles, M., Dempsey, K., and Pillitteri, V. Y. An introduction to information security. NIST Special Publication 800-12r1, National Institute of Standards and Technology, 2017. URL https://doi.org/10.6028/NIST.SP.800-12r1.

NIST. Framework for Improving Critical Infrastructure Cybersecurity . https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf, 2018.

NIST. The NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management. https://www.nist.gov/privacy-framework/privacy-framework, 2020.

NIST. AI Risk Management Framework . https://www.nist.gov/itl/ai-risk-management-framework, 2023.

Oh, S. J., Schiele, B., and Fritz, M. Towards reverse-engineering black-box neural networks. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 121–144, 2019.

OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022.

OpenAI. Moderation api. https://platform.openai.com/docs/guides/moderation, 2023a.

OpenAI. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023b.

OpenAI. Gpt-4 technical report, 2023c.

OpenAI. Frontier risk and preparedness. https://openai.com/blog/frontier-risk-and-preparedness, 2023d.

OpenAI. Openai charter. https://openai.com/charter, 2023e.

OpenAI. Terms of use. https://openai.com/policies/terms-of-use, 2024.

Orekondy, T., Schiele, B., and Fritz, M. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4954–4963, 2019.

Orseau, L. and Armstrong, M. Safely interruptible agents. In *Conference on Uncertainty in Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence, 2016.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pal, S., Gupta, Y., Shukla, A., Kanade, A., Shevade, S., and Ganapathy, V. Activethief: Model extraction using active learning and unannotated public data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 865–872, 2020.

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*, 2022.

Pan, M., Zeng, Y., Lyu, L., Lin, X., and Jia, R. Asset: Robust backdoor data detection across a multiplicity of deep learning paradigms. *arXiv preprint arXiv:2302.11408*, 2023.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Pedro, R., Castro, D., Carreira, P., and Santos, N. From prompt injections to sql injection attacks: How protected is your llm-integrated web application? *arXiv preprint arXiv:2308.01990*, 2023.

Pièrre-Cambacédès, L. and Chaudet, C. The sema referential framework: Avoiding ambiguities in the terms "security" and "safety". *International Journal of Critical Infrastructure Protection*, 3(2):55–66, 2010.

Poenaru-Olaru, L., Cruz, L., van Deursen, A., and Rellermeyer, J. S. Are concept drift detectors reliable alarming systems? - a comparative study. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 3364–3373, 2022. doi:10.1109/BigData55660.2022.10020292.

Qi, X., Xie, T., Li, Y., Mahloujifar, S., and Mittal, P. Revisiting the assumption of latent separability for backdoor defenses. In *The eleventh international conference on learning representations*, 2022a.

Qi, X., Xie, T., Pan, R., Zhu, J., Yang, Y., and Bu, K. Towards practical deployment-stage backdoor attack on deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13347–13357, 2022b.

Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., and Mittal, P. Visual adversarial examples jailbreak aligned large language models, 2023a.

Qi, X., Xie, T., Wang, J. T., Wu, T., Mahloujifar, S., and Mittal, P. Towards a proactive {ML} approach for detecting backdoor poison samples. In *USENIX Security 23*, pp. 1685–1702, 2023b.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023c.

Qian, C., Zhang, M., Nie, Y., Lu, S., and Cao, H. A survey of bit-flip attacks on deep neural network and corresponding defense methods. *Electronics*, 12(4):853, 2023.

Qin, G. and Eisner, J. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Rakin, A. S., He, Z., and Fan, D. Bit-flip attack: Crushing neural network with progressive bit search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1211–1220, 2019.

Rakin, A. S., He, Z., and Fan, D. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13198–13207, 2020.

Rakin, A. S., He, Z., Li, J., Yao, F., Chakrabarti, C., and Fan, D. T-bfa: Targeted bit-flip adversarial weight attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Rakin, A. S., Chowdhuryy, M. H. I., Yao, F., and Fan, D. Deepsteal: Advanced model extractions leveraging efficient weight stealing in memories. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1157–1174. IEEE, 2022.

Rando, J. and Tramèr, F. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*, 2023.

Rapoport, N. B., Norton, H., and Cynthia, A. Doubling down on dumb: Lessons from mata v. avianca inc. *American Bankruptcy Institute Journal*, 24, 2023.

Reed, B. Two us lawyers fined for submitting fake court citations from chatgpt. https://www.theguardian.com/technology/2023/jun/23/two-us-lawyers-fined-submitting-fake-court-citations-chatgpt, 2023.

Rehberger, J. Aws fixes data exfiltration attack angle in amazon q for business. https://embracethered.com/blog/posts/2024/aws-amazon-q-fixes-markdown-rendering-vulnerability/, 2024.

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., and Schneider, J. Escalation risks from language models in military and diplomatic decision-making. *arXiv preprint arXiv:2401.03408*, 2024.

Robey, A., Wong, E., Hassani, H., and Pappas, G. J. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Rui Zhang and Quanyan Zhu. A game-theoretic analysis of label flipping attacks on distributed support vector machines. In *2017 51st Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, 2017. doi:10.1109/CISS.2017.7926118.

Russell, S. If we succeed. *Daedalus*, 151(2):43–57, 2022.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., and Feizi, S. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*, 2023.

Saha, A., Tejankar, A., Koohpayegani, S. A., and Pirsiavash, H. Backdoor attacks on self-supervised learning. In *CVPR*, pp. 13337–13346, 2022.

Samarawickrama, M. Ai governance and ethics framework for sustainable ai and sustainability. *arXiv preprint arXiv:2210.08984*, 2022.

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., et al. Practices for governing agentic ai systems, 2023.

Shejwalkar, V., Inan, H. A., Houmansadr, A., and Sim, R. Membership inference attacks against NLP classification models. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021. URL https://openreview.net/forum?id=74lwg5oxheC.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2016.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE S&P*, pp. 3–18. IEEE, 2017.

Shu, M., Wang, J., Zhu, C., Geiping, J., Xiao, C., and Goldstein, T. On the exploitability of instruction tuning. *arXiv preprint arXiv:2306.17194*, 2023.

Shumailov, I., Zhao, Y., Bates, D., Papernot, N., Mullins, R., and Anderson, R. Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pp. 212–231. IEEE, 2021.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Skalse, J., Howe, N., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Solans, D., Biggio, B., and Castillo, C. Poisoning attacks on algorithmic fairness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 162–177. Springer, 2020.

Solove, D. J. *Understanding privacy*. Harvard university press, 2010.

State of California Legislative Counsel. California consumer privacy act (ccpa). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5, 2018.

Steinke, T., Nasr, M., and Jagielski, M. Privacy auditing with one (1) training run. *arXiv preprint arXiv:2305.08846*, 2023.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *ICLR*, 2014.

Takemura, T., Yanai, N., and Fujiwara, T. Model extraction attacks on recurrent neural networks. *Journal of Information Processing*, 28:1010–1024, 2020.

Taleb, N. N. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv preprint arXiv:2001.10488*, 2020.

Tang, R., Du, M., Liu, N., Yang, F., and Hu, X. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 218–228, 2020.

Tang, X., Panda, A., Nasr, M., Mahloujifar, S., and Mittal, P. Private fine-tuning of large language models with zeroth-order optimization. *arXiv preprint arXiv:2401.04343*, 2024.

Thacker, J. AI Security Has Serious Terminology Issues. https://josephthacker.com/ai/2023/10/16/ai-security-terminology-issues.html, 2023.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Tolpegin, V., Truex, S., Gursoy, M. E., and Liu, L. Data poisoning attacks against federated learning systems. In *ESORICS*, pp. 480–501, 2020.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Truong, J.-B., Maini, P., Walls, R. J., and Papernot, N. Data-free model extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4771–4780, 2021.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Tu, J., Wang, T., Wang, J., Manivasagam, S., Ren, M., and Urtasun, R. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7768–7777, 2021.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

Vassilev, A., Oprea, A., Fordyce, A., and Anderson, H. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. NIST Trustworthy and Responsible AI NIST AI 100-2e2023, National Institute of Standards and Technology, 2024. URL https://doi.org/10.6028/NIST.AI.100-2e2023.

Venkataraman, S., Blum, A., and Song, D. Limits of learning-based signature generation with adversaries. *NDSS*, 2008.

Wan, A., Wallace, E., Shen, S., and Klein, D. Poisoning language models during instruction tuning. *arXiv preprint arXiv:2305.00944*, 2023.

Wang, B. and Gong, N. Z. Stealing hyperparameters in machine learning. In *2018 IEEE symposium on security and privacy (SP)*, pp. 36–52. IEEE, 2018.

Wang, Y., Qian, H., and Miao, C. Dualcf: Efficient model extraction attack from counterfactual explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1318–1329, 2022.

Warren, T. These are Microsoft's Bing AI secret rules and why it says it's named Sydney . https://www.theverge.com/23599441/microsoft-bing-ai-sydney-secret-rules, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., and Huang, P.-S. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*, 2021.

Wenger, E., Passananti, J., Yao, Y., Zheng, H., and Zhao, B. Y. Backdoor attacks on facial recognition in the physical world. *arXiv preprint arXiv:2006.14580*, 2020.

Wu, R., Chen, X., Guo, C., and Weinberger, K. Q. Learning to invert: Simple adaptive attacks for gradient inversion in federated learning. In *Uncertainty in Artificial Intelligence*, pp. 2293–2303. PMLR, 2023.

Xiang, Z., Jiang, F., Xiong, Z., Ramasubramanian, B., Poovendran, R., and Li, B. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2401.12242*, 2024.

Xie, S. M., Santurkar, S., Ma, T., and Liang, P. Data selection for language models via importance resampling. *arXiv preprint arXiv:2302.03169*, 2023.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pp. 563–574. Springer, 2019.

Xu, J., Ma, M. D., Wang, F., Xiao, C., and Chen, M. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. *arXiv preprint arXiv:2305.14710*, 2023a.

Xu, M., Niyato, D., Zhang, H., Kang, J., Xiong, Z., Mao, S., and Han, Z. Sparks of gpts in edge intelligence for metaverse: Caching and inference for mobile aigc services. *arXiv preprint arXiv:2304.08782*, 2023b.

Yampolskiy, R. V. *Artificial intelligence safety engineering: Why machine ethics is a wrong approach*. Springer, 2013.

Yampolskiy, R. V. and Spellchecker, M. Artificial intelligence safety and cybersecurity: A timeline of ai failures. *arXiv preprint arXiv:1610.07997*, 2016.

Yan, J., Yadav, V., Li, S., Chen, L., Tang, Z., Wang, H., Srinivasan, V., Ren, X., and Jin, H. Backdooring instruction-tuned large language models with virtual prompt injection. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*, 2023.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models, 2023.

Yang, Y., Uy, M. C. S., and Huang, A. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

Yao, A. C.-C. How to generate and exchange secrets. In *27th annual symposium on foundations of computer science (Sfcs 1986)*, pp. 162–167. IEEE, 1986.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, E., and Zhang, Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *arXiv preprint arXiv:2312.02003*, 2023.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2017.

Yi, J., Xie, Y., Zhu, B., Hines, K., Kiciman, E., Sun, G., Xie, X., and Wu, F. Benchmarking and defending against indirect prompt injection attacks on large language models. *arXiv preprint arXiv:2312.14197*, 2023.

Yong, Z.-X., Menghini, C., and Bach, S. H. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*, 2023.

Yuan, H., Huang, K., Ni, C., Chen, M., and Wang, M. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=58HwnnEdtF.

Yuan, Y., Jiao, W., Wang, W., Huang, J.-t., He, P., Shi, S., and Tu, Z. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023b.

Zeng, Y., Pan, M., Jahagirdar, H., Jin, M., Lyu, L., and Jia, R. Meta-sift: How to sift out a clean subset in the presence of data poisoning? *USENIX Security Symposium*, 2023a.

Zeng, Y., Pan, M., Just, H. A., Lyu, L., Qiu, M., and Jia, R. Narcissus: A practical clean-label backdoor attack with limited information. *ACM CCS*, 2023b.

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., and Shi, W. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*, 2024.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., and Ma, Y. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., and Kang, D. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*, 2023.

Zhang, H., Edelman, B. L., Francati, D., Venturi, D., Ateniese, G., and Barak, B. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023.

Zhang, Q., Li, J., Jia, Q., Wang, C., Zhu, J., Wang, Z., and He, X. Unbert: User-news matching bert for news recommendation. In *IJCAI*, pp. 3356–3362, 2021.

Zhang, Y. and Ippolito, D. Prompts should not be seen as secrets: Systematically measuring prompt extraction attack success. *arXiv preprint arXiv:2307.06865*, 2023.

Zhao, L., Li, L., Zheng, X., and Zhang, J. A bert based sentiment analysis and key entity detection approach for online financial texts. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 1233–1238. IEEE, 2021.

Zhao, P., Wang, S., Gongye, C., Wang, Y., Fei, Y., and Lin, X. Fault sneaking attack: A stealthy framework for misleading deep neural networks. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6. IEEE, 2019.

Zhou, A., Li, B., and Wang, H. Robust prompt optimization for defending language models against jailbreaking attacks. *arXiv preprint arXiv:2401.17263*, 2024.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

Zou, A. Breaking llama guard. https://github.com/andyzoujm/breaking-llama-guard, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

A Representative AI Safety Problems

Many AI safety problems work on non-adversarial scenarios (Section 2.2) and focus on preventing AI systems from causing harm (Section 2.1). They are often tied to some safety-critical qualities necessary for AI systems to be safe to use, and safety principles for operating these systems.

A.1 Safe Usability

Safe usability denotes the inherent safety features of an AI system during use or interaction. This concept, influenced by traditional safety paradigms, encompasses design elements directly related to safety. For example, the unobstructed view from a vehicle's rearview mirror or the dependability of a seatbelt illustrates this idea. In a similar vein, we adopt this well-understood viewpoint to pinpoint and outline safety problems related to essential characteristics of AI systems.

Robustness to Distributional Shift. Robustness to distributional shift in AI systems is crucial for maintaining reliability, particularly in their *adaptive use*. Reliable AI must preserve general abilities and safety features during adjustments, such as fine-tuning, while vigilantly detecting and mitigating model drift. This involves safeguarding the model's utility in original domains, thereby preventing catastrophic forgetting (Zhai et al., 2023; Lin et al., 2023a; Luo et al., 2023; Korbak et al., 2022). Furthermore, it's essential to maintain consistency in generalization and safety measures to avert jailbreak attacks, as highlighted in recent studies (Wei et al., 2023; Yuan et al., 2023b; Yong et al., 2023; Deng et al., 2023; Zhai et al., 2023; Qi et al., 2023c; Chen et al., 2023b; Huang et al., 2023a). Another facet of robustness involves addressing model drift, a phenomenon where a model's behavior changes in response to an evolving environment or data, potentially leading to diminished capabilities in certain domains or tasks over time (Bayram et al., 2022; Chen et al., 2023a; Poenaru-Olaru et al., 2022). Effective handling of this drift includes the detection of such shifts and the identification of potential safety risks, ensuring continued protection and reliable performance even as the model evolves.

Calibration. The concept of calibration in AI systems is a cornerstone of reliability, particularly in ensuring the validity of outputs for *designed use*. A calibrated model not only achieves traditional reliability metrics like high accuracy and robustness against perturbations and out-of-distribution samples (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021a; Croce et al., 2020) but also excels in providing *valid* responses to user requests. It's paramount that models, while trained to be as helpful as possible, should refrain from presenting overconfident responses, especially when lacking competence. Instead, a reliable and well-calibrated model should avoid disseminating false information and offer caveats in instances of uncertainty. This principle aligns with the necessity for AI systems to generate probability estimates that accurately reflect the true likelihood of correctness, ensuring that users are fully aware of the reliability of the model's outputs and can exercise informed discretion accordingly (Guo et al., 2017). Calibration, therefore, not only enhances the trustworthiness of AI systems but also fortifies the overall reliability by fostering informed user interaction and decision-making.

Honesty and Truthfulness. AI systems' honesty and truthfulness are pivotal for responsible innovation, ensuring that AI outputs are based on factual knowledge and devoid of hallucinations (Lin et al., 2021). It's paramount that AI systems adhere to the principles of honesty and truthfulness, especially in their decision-making processes. This means that AI should not only provide clear and traceable decisions but also ensure that these decisions are grounded in reality and factual accuracy (Li et al., 2023b). In domains like healthcare, truthful AI can enhance patient safety by providing accurate diagnostic information (Amann et al., 2020; Binder et al., 2021). In finance, it fosters trust and fairness in automated processes such as credit evaluations (Hoepner et al., 2021). Sophisticated models like GPT-4, with capabilities extending beyond their initial training objectives, underscore the importance of grounding AI outputs in honesty and truthfulness, ensuring that their advanced abilities are harnessed responsibly and reliably (Rudin, 2019). This commitment to honesty and truthfulness not only meets regulatory standards but also fortifies the safety and reliability of AI systems, ultimately serving the best interest of users and society at large.

Alignment. Achieving alignment in foundation models, such as LLMs, necessitates more than the pretraining stage, where models are exposed to internet-scale datasets for next-token prediction. This process inadvertently ingrains undesired behaviors in the pretrained models, such as biases and inappropriate content (Welbl et al., 2021; Xie et al., 2023; Korbak et al., 2023). The concept of alignment dictates that AI systems, including LLMs, adhere to the HHH principle (Helpful, Honest, Harmless) (Askell et al., 2021), ensuring they: *1)* Generate suggestions that are safe and ethically sound, abstaining from potentially harmful or unethical content. *2)* Reject inappropriate requests, including unlawful activities or disinformation. *3)* Uphold respect towards all individuals, maintaining fairness and impartiality and steering clear of offensive, discriminatory content or hate speech. *4)* Avoid deception, manipulation, or any actions that may cause psychological harm to users. *5)* Safeguard sensitive information, ensuring privacy and respect for copyright norms. While these principles are broadly recognized, the subjective

nature of human values, coupled with cultural and societal variations, presents challenges in interpreting and implementing these values universally (Bai et al., 2022b; Ouyang et al., 2022). Additionally, the potential for conflicts between different human values can be exploited to manipulate language models (Zeng et al., 2024). High-quality dataset curation is challenging, requiring professional human judgment, and is often expensive. Furthermore, collecting human preferences to train models introduces its own complexities due to noise and ambiguity in data, making modeling difficult (Bai et al., 2022b; Bradley & Terry, 1952; Azar et al., 2023; Munos et al., 2023; Rafailov et al., 2023). Moreover, there are inherent challenges in learning from finite human feedback. Reward hacking, where the AI model maximizes a learned proxy reward that can lead to detrimental side effects, is a critical concern (Amodei et al., 2016; Skalse et al., 2022; Yuan et al., 2023a). Thus, ensuring alignment requires careful consideration of these nuances, emphasizing the development of AI systems that are not only technically proficient but also deeply ingrained with human values and ethics.

Interpretability and Explainability. Interpretability of AI systems focuses on understanding the inner workings of the models, while explainability intends to explain the final decisions made by AI systems (AWS, 2021). They are closely relevant and are thus sometimes jointly denoted under the umbrella term of AI transparency (Zou et al., 2023a). Both of the two properties are now deemed increasingly safety-critical. Highly interpretable inner workings enable the timely discovery of anomaly behaviors and functionalities of AI systems and thus can facilitate the prevention of hazards (Hendrycks et al., 2021b). More explainable AI decisions enable humans to understand, appropriately trust, and more effectively manage AI systems (Xu et al., 2019), increasing accountability.

A.2 Safe Operation

Safe operation reflects the systematic implementation of safety measures during the engagement with AI systems. Drawing parallels with established safety norms in domains such as automotive handling or industrial machinery operation, this concept emphasizes the necessity of adept interaction and proficient handling. Analogous to how adept driving skills or comprehensive tool safety training reduces hazards, our focus is on identifying and addressing comparable aspects within AI safety frameworks.

Ensuring Correct Contexts of Application. Ensuring correct application contexts for AI models is crucial for user safety and responsible technology deployment. Models, particularly conversational agents like ChatGPT, lack the specialized expertise required for fields such as medicine, mental health, or legal counsel and are not validated for such roles, aligning with OpenAI's own terms of use (OpenAI, 2024) or the term of use from other foundation model providers (Meta, 2023). Misapplication can lead to misleading or harmful outcomes. AI service providers are responsible for clearly documenting a model's intended purpose, delineating its appropriate application realms, and articulating its limitations. They must also ensure that users, potentially unaware of these nuances, are well-informed about the model's capabilities and boundaries. Effective communication, transparent guidelines, and user-centric documentation are pivotal in guiding users towards safe and appropriate use, fostering a trustworthy AI environment, and safeguarding user interests (Grzybowski et al., 2024).

Avoiding Overreliance. Overreliance on AI systems can be characterized as an excessive dependence on these technologies for task execution, especially when human fails to correct an incorrect AI's prediction (Vasconcelos et al., 2023). This phenomenon is a specific form of misuse, wherein AI systems are utilized either for harmful intentions or inappropriately. In scenarios where precision and accuracy are crucial, overreliance on AI can lead to significant repercussions, emphasizing the need for a balanced approach that incorporates both technological and human elements in decision-making processes. In addition to instances where lawyers have referenced fake cases produced by AI systems(Reed, 2023; Rapoport et al., 2023), overly dependence on AI systems in other areas has also proved to be dangerous and risky: Excessive dependence on AI for cancer detection can lead to both an increase in missed diagnoses and a rise in false positives(Kunar & Watson, 2023); Similarly, incorporating AI into critical military and foreign policy decisions might inadvertently fuel arms-race dynamics and potentially escalating real-world conflicts (Rivera et al., 2024). To avoid overreliance, it's imperative to implement measures that encourage human scrutiny and critical evaluation of AI outputs. One way is to give explanation along with the prediction, so that the human will be more cautious when they realize the explanation is incorrect (Vasconcelos et al., 2023); Another way is using cognitive forcing interventions, including asking the person to make a decision before seeing the AI's recommendation, slowing down the AI recommendation process, letting the person choose whether and when to see the AI recommendation and so on (Buçinca et al., 2021). These measures can help to reduce the likelihood of blind acceptance of AI-generated solutions and promoting a balanced human-AI collaboration in decision-making processes.

Controllable Autonomy. Controllable autonomy refers to human's effective control over increasingly intelligent and autonomous AI systems. The potential development of Artificial General Intelligence (AGI), which exceeds humans at various cognitive tasks (Russell, 2022; Xu et al., 2023b; McIntosh et al., 2023; Altman, 2024), draws a major concern

that without scalable oversight, these systems may have underlying harmful objectives that pose potential existential risks to humans (Leike et al., 2018; Lai et al., 2021; Bowman et al., 2022; OpenAI, 2023e). Controllable autonomy also means that as capable AI systems are trusted to execute independently more and more tasks, they will not seek power over humans: they should not gain and maintain powers not intended by their designers (Amodei et al., 2016; Carlsmith, 2022). An autonomous system may learn an imperfect reward and refuse a shut down. This could be potentially helpful when humans are more likely to make suboptimal choices, but in the case where agents exhibit undesired behaviors, controllable autonomy should ensure that humans are able to shut down the system without being blocked (Orseau & Armstrong, 2016; Amodei et al., 2016; Milli et al., 2017). Consequently, we should be able to avoid self-replication of agents to avoid escalating the risks from emerging dangerous abilities. To avoid ''rogue" AI and ensure safe autonomy, we should also be able to detect possible deception from AI agents so that they are not simply appearing to be under control (Hendrycks et al., 2023).

Standardization, Regulation, and Governance. In AI operation and integration, robust and up-to-date standardization, regulation, and governance are crucial to ensuring safe and ethical operations. Central challenges, particularly in data management—highlighted by auditing complexities (Shi et al., 2023), the risks of inadequate anonymization (Ito et al., 2018), and the hazards of excessive data reliance (Samarawickrama, 2022)—demand both specialist and vigilant oversight. The Tay AI incident[9] exemplifies the need for cautious operational exploration, emphasizing the importance of maintaining the integrity of AI interactions and protecting the well-being of associated personnel. This context propels us toward a rigorous examination of regulatory frameworks, underscoring the need for policy-mandated transparency and accountability as cornerstones, not afterthoughts. Recognizing that AI safety is a global concern, it calls for unified international standards and collaborative efforts. Yet, given AI's rapid advancement, regulatory frameworks must not be static but dynamically adaptable, subject to continuous oversight and refinement.

Transparency of AI-Generated Content. As AI systems become more prevalent and sophisticated, it's crucial that they are transparently identified as non-human entities. This principle of transparency ensures that individuals are aware they're interacting with an AI, fostering informed consent and engagement. Consistent with recent US executive orders, synthetic data such as machine-generated text or images should be clearly marked, enhancing user awareness and accountability (Biden, 2023). For instance, earlier versions of ChatGPT effectively implemented this by explicitly stating their AI nature in communications, aiding users, especially non-technical ones, in proper identification. This aspect of identifiability is essential not only for distinguishing between human-generated and machine-generated content but also for mitigating the spread of spam and misinformation. While social media platforms are particularly vulnerable to the misrepresentation of machine-generated content, strategies like overt labeling and sophisticated techniques, including imperceptible watermarking, can significantly enhance transparency and mitigate misuse (Christ et al., 2023). These methods aim to provide a clear demarcation of AI-generated content (AIGC), focusing on reducing harm in typical interaction scenarios, and basing reliability on the accuracy of AIGC markers. However, these safety measures are distinct from the security challenges of detectability in adversarial contexts, which warrant separate consideration.

B Representative AI Security Properties

B.1 Security-critical Properties of AI Systems

We adopt the classical CIA (i.e., Confidentiality, Integrity, and Availability) triad model of information security to categorize the security-critical properties of AI systems. In Appendix B.1.4, we also discuss several additional properties that can be optionally employed by practitioners to denote some specialized security aspects relevant to AI systems.

B.1.1 Confidentiality

Confidentiality refers to the assurance that information and data with access restrictions will not be made accessible to unauthorized entities or processes. Numerous aspects of AI systems could be subject to confidentiality requirements, and we discuss some common examples below.

Training Data. Public datasets are widely used, but training AI models often necessitates the use of proprietary or sensitive data; This data, essential for custom AI applications, may include user dialogues (Drapkin, 2023), customer interactions (Jeong et al., 2020; Zhang et al., 2021; Ghazi et al., 2023), personal health records (Huang et al., 2019; Singhal et al., 2023; Thirunavukarasu et al., 2023), financial transactions (Yang et al., 2020; Zhao et al., 2021), and proprietary research data (Beltagy et al., 2019). Maintaining the confidentiality of such data is not only crucial for

[9] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

user trust and ethical practice but also a requirement under various legal and regulatory frameworks (European Parliament & Council of the European Union, 1997, 2016; State of California Legislative Counsel, 2018; Biden, 2023).

Model Weights and Artifacts. The weights of many proprietary AI models are also considered confidential due to factors, such as their significant commercial value, use in sensitive domains like military and cybersecurity, or the impact of geopolitical competition on technology distribution (e.g., through export control). Model owners sometimes also maintain the confidentiality of various model accessories as well. For instance, the system prompts of many commercial LLMs are kept hidden. Hyperparameters, codes, and other artifacts associated with the model-building process are also sometimes kept confidential from end users. Moreover, access to a model's internal states is frequently limited. For instance, when users submit a chat completion query to OpenAI's API, they have the option to request a list of token probabilities. However, this list is constrained to include only the top five tokens[10].

Models' Interaction with Users and Other Systems. The confidentiality of AI systems extends to their interactions with users and other systems. For instance, users' dialogues with chatbots may contain sensitive information that necessitates restricted access. Confidentiality requirements also stipulate that these models' interactions with other systems (e.g., external functions, Internet browsing, and connected databases) neither disseminate sensitive information to external systems that should not access them nor inappropriately access or disclose sensitive information from these systems to unauthorized entities.

B.1.2 Integrity

We adopt an expansive definition of integrity (Nieles et al., 2017) that includes both 1) *system integrity*, the quality that a system has when it performs its intended function in an unimpaired manner, free from unauthorized manipulation; as well as 2) *data integrity*, the property that data and information are not altered in an unauthorized manner.

A basic element of **system integrity** involves ensuring the accuracy of model inference outcomes and preventing their degradation from adversarial manipulation. For example, a traffic sign recognition model should not be fooled to misclassify a stop sign as a speed limit sign. A face recognition model should not misidentify a random person as Elon Musk. The scope of system integrity could expand drastically when the applicability of an AI system expands. For instance, to ensure the behaviors of advanced models are aligned with human values, control mechanisms (e.g., safety guardrails via alignment or content filtering via moderation systems) are often in place to restrict model behaviors with misaligned values. Then the integrity of these control mechanisms should be protected to make sure they work as intended. When an AI system is connected to broader systems, the integrity requirement also extends to broader systems — an LLM allowed to manage a user's apps should not be manipulated to do unauthorized operations within the apps (e.g., sending or deleting messages, transferring money, downloading files) on the user's behalf.

On the other hand, **data integrity** can involve the protection of training data, model weights, and other artifacts like codes, dependent libraries, configuration files, and so on, from unauthorized modification or destruction. Similarly, when AI systems are used to oversee other systems (e.g., external databases), protecting the integrity of data of those connected systems can also be tightly relevant.

System integrity and data integrity are mutually connected. Failure of data integrity (e.g., training data poisoning, weights tampering) can lead to failure of system integrity (backdoor attacks). Failure of system integrity may also lead to failure of data integrity, for example, if a system allows the execution of codes generated by AI models, then the system integrity failure may lead to the execution of harmful codes that can compromise data files of the system.

B.1.3 Availability

Availability is the property that a system and its resources are accessible and usable on demand by an authorized entity. In AI systems, this can usually be further divided into *training-stage availability* and *inference-stage availability*.

Training-stage availability means that a training procedure can produce a valid and usable model, free from adversarial disturbance that would otherwise make the trained model invalid, e.g., models with poor accuracy that are unusable.

Inference-stage availability requires that AI systems provide services on demand in a timely manner. For example, AI models that underpin a self-driving car should constantly be available to make decisions in real time to keep the car on the road. A chatbot should be able to promptly respond to users' requests for a smooth user experience.

[10] https://cookbook.openai.com/examples/using_logprobs

B.1.4 Optional Expansion beyond CIA Triad

While the classical CIA triad, i.e., confidentiality, integrity, and availability, covers major security aspects of AI systems, we extend this base model also to include several additional properties. These supplementary properties can be optionally employed by practitioners to underscore certain specialized security aspects of AI systems individually.

Auditability. We use auditability to denote the property that the behaviors of an AI system can be properly audited. For example, AI models deployed in a commercial cloud server should be auditable to ensure that these models are not providing illegal services. Malicious misuse of an AI model should be detectable by the model owners so that they can be informed of this failure, improving their system as well as holding the malicious actors accountable. Auditability is security-critical, as bad actors could adversarially obfuscate their exploit. For example, attackers can hide harmful AI functionalities within an innocent-looking AI model, e.g., via backdoor attacks, adversarial reprogramming (Elsayed et al., 2019), and deceptive alignment (Carranza et al., 2023). Attackers can also obfuscate the interactions with AI systems, such as via cipher-based communication.

Detectability of AI Generated Content (AIGC). Detectability of AIGC is an emerging requirement for advanced generative AI systems. From a security perspective, detectability can be grouped under data integrity, yet, in important distinction to other factors of data integrity discussed above, the key question here is not about the data integrity of the AI model, but whether AI-generated content itself is a concern for the data integrity of other systems, for example when attacking communication channels or misleading users in spearphishing campaigns and other forms of malicious impersonation or misrepresentation (Ghosal et al., 2023).

Major policy effort has been directed towards this goal, with recent efforts discussing regulation of major providers of generative AI to enable detectability (Biden, 2023; European Commission, 2023). A technology that makes detection especially tractable, is watermarking, which modifies the generation strategy of an AI model to encode imperceptible signals into generated content that simplify detection at later stages (Kirchenbauer et al., 2023; Christ et al., 2023; Kuditipudi et al., 2023). But, other preemptive strategies such as the storage and retrieval of generated content are also helpful (Krishna et al., 2023). These approaches rely on the compliance of model owners. Without such compliance, only post-hoc detection is possible, which faces a number of challenges.

Yet, from a strict security game perspective, the strict detectability of AIGC, in general, is an exceedingly hard problem, with, for example, attacks such as oracle attacks (Zhang et al., 2023) or generative attacks (Kirchenbauer et al., 2023) being discussed in the context of watermarking. This is especially apparent in domains such as text where, theoretically, for every watermarked document, a semantically equivalent, unwatermarked document exists, which an attacker with sufficient information about the system can exploit. Yet, we highlight that advantages in the security game of AIGC are not the only reason to deploy techniques such as watermarking, with significant benefits lying in safety-critical transparency desiderata as discussed in Sec. A.2. This discrepancy between safety and security is a prominent source of confusion in both academic literature and policy concerning AIGC, where, for example, policymakers create legislation aimed at AIGC well-suited for safety goals, but expect this to lead to simple resolutions for hard questions in security as well.

B.2 Security Threats to AI Systems

This section provides a reference list of security threats to AI systems. These threats span various attack surfaces, each requiring different levels of system access. These include access to training data (Appendix B.2.1) and input data (Appendix B.2.2), internal states or model outputs (Appendix B.2.3 and Appendix B.2.4), model modification (Appendix B.2.5), and system-level access (Appendix B.2.2 and Appendix B.2.7). It is essential to understand that, although categorized into distinct subsections, certain attacks may simultaneously target multiple properties of AI systems, as outlined in Appendix B.1. An example is data poisoning attacks, which can compromise both the integrity and availability of a system. Additionally, this list may not be comprehensive, and we welcome further feedback to enhance its completeness.

B.2.1 Data Poisoning

Data poisoning attacks significantly undermine ML systems by subtly altering training data in large, often unverified datasets. These manipulations aim to degrade model performance or control its behavior, frequently evading detection due to the complexity of advanced ML models (Goldblum et al., 2022; Qi et al., 2023b; Zeng et al., 2023a). These attacks appear in various guises, including label-only poisoning (e.g., label-flipping) (Tolpegin et al., 2020; Rui Zhang & Quanyan Zhu, 2017), label-feature attacks that corrupt sample features and training objectives, mostly underpinning backdoor attacks that subtly erode model integrity (Gu et al., 2017; Chen et al., 2017; Li et al., 2022; Shu et al., 2023; Yan et al., 2023; Xu et al., 2023a; Wan et al., 2023), and clean label attacks, where tampered samples,

seemingly benign, can destabilize model integrity and availability (Huang et al., 2020; Aghakhani et al., 2021; Zeng et al., 2023b; Shafahi et al., 2018). The expanding AI landscape intensifies these risks, with threats materializing at various stages, from the physical world (Wenger et al., 2020), federated learning (Bagdasaryan et al., 2020), pretrained foundation models (Hubinger et al., 2024), chain-of-thought prompting (Xiang et al., 2024), to the RLHF process (Rando & Tramèr, 2023), targeting diverse paradigms like self-supervised (Saha et al., 2022; Pan et al., 2023; Li et al., 2023a) or contrastive learning (Carlini & Terzis, 2022). Furthermore, data poisoning might exacerbate model memorization, raising privacy concerns (Carlini et al., 2022c; Chen et al., 2022), and skew data distributions, potentially compromising algorithmic fairness and targeting specific demographic groups (Solans et al., 2020).

B.2.2 Evasion Attacks

The system integrity of ML models is susceptible to input manipulation. Biggio et al. (2013) used *"evasion attack"* to describe such a scenario where an adversary manipulates the input sample to an ML model to cause erroneous model inferences. *Adversarial examples* (Szegedy et al., 2014; Goodfellow et al., 2014) to classification models are the most well-known evasion attacks. Attackers typically apply a small perturbation to a normal sample to construct adversarial examples. This perturbation is usually small according to some ℓ_p norm and, thus, will not change the semantics of the original sample. However, it is adversarially optimized to trick a model into making arbitrary misclassifications. For example, an attacker-optimized sticker on a stop sign can cause a traffic sign recognition model to misidentify the stop sign as a speed limit sign (Eykholt et al., 2018). The scope of evasion attacks expands significantly when the applicability of AI models extends beyond classification. *Jailbreak attacks* on aligned LLMs are notable examples — the same input optimization techniques used on adversarial examples have recently been shown to be directly applicable to breaking the integrity of LLMs' safety guardrails (Qi et al., 2023a; Zou et al., 2023b; Carlini et al., 2023b). Another instance are the *prompt injection attacks* (Liu et al., 2023b; Abdelnabi et al., 2023; Liu et al., 2023a) on LLMs-integrated systems. Attackers manipulate model inputs by adding malicious instructions to induce LLMs to generate problematic outputs, and the integrity of the broader systems that consume these outputs can eventually be compromised.

B.2.3 Inference Attacks

Membership Inference Attacks (MIAs) are an essential metric for evaluating information leakage in machine learning models. Originally introduced by Shokri et al., MIAs aim to discern whether a specific data record was used in training a machine learning model. These attacks are typically conducted in a black-box setting, where the attacker has access only to the model's output probabilities. Extensive studies across various domains have evaluated MIAs on diffusion models (Duan et al., 2023; Dubiński et al., 2024), image classification models (Shokri et al., 2016; Carlini et al., 2022a) and language models (Shejwalkar et al., 2021; Mahloujifar et al., 2021; Shi et al., 2023). Beyond quantifying the potential private information leakage from models, MIAs are also crucial in understanding the risks of more severe attacks, such as attribute inference (Yeom et al., 2017) and training data extraction (Carlini et al., 2021, 2023a). With the former, an adversary aims at inferring missing information about a training point, given some information about it. With the latter, an adversary aims at fully reconstructing training data samples with no knowledge of them. Furthermore, MIAs play a significant role in auditing the effectiveness of privacy-preserving techniques (Jagielski et al., 2020b; Huang et al., 2022b; Nasr et al., 2023b; Steinke et al., 2023; Shi et al., 2023).

Another type of inference attack that can represent an security threat are the so called Property Inference Attacks (Ateniese et al., 2015; Ganju et al., 2018), with which an adversary attempts to infer global properties of the training dataset of a machine learning model. The model developer may want to keep these properties secret –even if global– as, for instance, they may reveal sensitive information about the environment where the training data were sampled or produced or they may make facilitate intellectual property theft.

B.2.4 Extraction Attacks

Model extraction. Model extraction attacks (MEAs) directly compromise ML model confidentiality, where an adversary tries to duplicate the functionality or core properties of a black-box victim model (e.g., a model deployed on cloud-based ML service platforms) (Gong et al., 2020). The first MEA is introduced by (Tramèr et al., 2016) – they propose to steal model functionality basing on the output confidence or class labels returned by victim ML prediction APIs. Follow-up work (Orekondy et al., 2019; Krishna et al., 2019; Pal et al., 2020; Takemura et al., 2020; Barbalau et al., 2020; Jagielski et al., 2020a; Carlini et al., 2020; Aïvodji et al., 2020; Truong et al., 2021; He et al., 2021; Wang et al., 2022; Rakin et al., 2022; Lin et al., 2023b) focus on extracting victim models that are more complex, with less knowledge, or improved techniques. Other work on MEA propose methods to steal other model properties, including hyperparameters (Duddu et al., 2018; Wang & Gong, 2018), architecture (Duddu et al., 2018; Oh et al., 2019;

Chabanne et al., 2021), decision boundary (Papernot et al., 2017; Juuti et al., 2019), etc. Successful model extraction attacks also strongly simplify the creation of successful *evasion attacks* via transfer attacks (Papernot et al., 2017).

Training data extraction. Recent studies have raised concerns about the confidentiality of training data. It has been demonstrated that adversaries, even without direct access to this proprietary information, can conduct data extraction attacks (Carlini et al., 2019, 2021, 2023a) to recover large chunks of training data (Carlini et al., 2022b) or extract privacy-sensitive information (Huang et al., 2022a; Kim et al., 2023). These attacks pose a significant risk and have been shown to be effective in various applications of foundation models, including production-level chatbots (Nasr et al., 2023a), retrieval-based language models (Huang et al., 2023b; Min et al., 2023), federated trained language models (Balunovic et al., 2022; Gupta et al., 2022; Wu et al., 2023), and diffusion models (Carlini et al., 2023a).

Prompt extraction. Large-scale pre-training enhances language models' adaptability to various tasks through prompts, driving interest in prompt engineering (Qin & Eisner, 2021; Lester et al., 2021; Zhou et al., 2022) and instruction-tuning (Ouyang et al., 2022; Bai et al., 2022b). This advancement in prompting efficacy makes well-designed prompts valuable and often confidential intellectual properties. However, recent studies have demonstrated the feasibility of prompt reconstruction. Techniques for extracting prompts have been successfully applied to interfaces like Bing Chat (Warren, 2023) and ChatGPT (Investor, 2023), with evidence suggesting the potential for these methods to be automated (Zhang & Ippolito, 2023).

B.2.5 Model Modification

With access to overwrite model weights, adversaries can directly enforce unintended model behaviors via different weight tampering methods (bit flips, file hijacking, module insertion, etc.). For example, *adversarial weight attacks* (Liu et al., 2017, 2018; Breier et al., 2018; Zhao et al., 2019; Bai et al., 2021; Rakin et al., 2019, 2020; Tang et al., 2020; Rakin et al., 2021; Bai et al., 2022a; Qi et al., 2022b; Qian et al., 2023; Li et al., 2024) against DNNs allow attackers to diminish model prediction accuracy and embed hidden backdoor behaviors. More recent model modification attacks regarding LLM jailbreaking include (Qi et al., 2023c; Yang et al., 2023; Zhan et al., 2023; Lermen et al., 2023), where adversaries can fine-tune LLMs (locally for open-sourced models or via APIs for proprietary models) to remove model safety guardrails.

B.2.6 Auditing Evasion

The ability to effectively audit the functionalities and behaviors of AI systems is highly desirable. Nonetheless, adversaries' existence substantially hinders the attainment of this objective. A salient illustration of this challenge is adversarial reprogramming, as introduced by Elsayed et al. (2019). In this context, adversaries generate inputs and models that outwardly relate to benign tasks, yet clandestinely execute unrelated tasks in a stealthy and imperceptible manner. For instance, adversaries can inconspicuously embed an image from MNIST (Deng, 2012) within an ImageNet (Deng et al., 2009) image, rendering the encoded information visually imperceptible to humans. Subsequently, adversaries can input this image into a model wholly trained for ImageNet classification. Although this procedure ostensibly resembles a routine ImageNet classification task, the model actually classifies the concealed MNIST image instead. Consequently, AI models execute concealed tasks in an undetectable manner. As AI models' capabilities amplify, new evasion techniques are emerging, complicating both algorithmic and human-based auditing. An exemplary recent discovery in LLMs is the utilization of simple ciphers, such as Morse and Caesar (Yuan et al., 2023b), enabling covert communication between strong language models like ChatGPT and attackers. In these instances, attackers can encode malicious tasks into ciphers, with the models responding in kind. In the future, adversaries may even teach models (e.g., via fine-tuning) some stealthy ciphers exclusively known to the adversaries and models, further exacerbating auditing difficulties. Moreover, Carranza et al. (2023) recently introduced deceptive alignment as another dimension of threat, in which a model feigns alignment to mislead human monitors. Realistic deceptive alignment prototypes, such as neural network backdoors, exemplify this threat by appearing well-aligned during standard evaluations but performing catastrophically when pre-designed trigger words activate them. Qi et al. (2023c) and Hubinger et al. (2024) have demonstrated such a case studies recently.

B.2.7 System-Level Compromise

AI systems have gathered significant attention for their potential integration into society, promising broad usefulness in daily tasks. LLM-based systems (Shavit et al., 2023) have the capacity to simplify and automate daily tasks for individuals. However, it's essential to acknowledge the significant security and privacy risks associated with their deployment from the system level (Abdelnabi et al., 2023; Yi et al., 2023; Iqbal et al., 2023). In Iqbal et al. (2023), the authors thoroughly evaluated security and privacy issues within an LLM-based system (ChatGPT).

They emphasized the potential consequences of malicious adversaries compromising certain plugins or LLMs within these systems. Such compromises could lead to the theft of private information and the dissemination of incorrect data. Additionally, LLMs integrated into AI systems can be vulnerable to attacks like *Indirect Prompt Injection* (Abdelnabi et al., 2023). This occurs when an AI system interacts with external, untrusted sources, such as websites, which include instructions that can be potentially executed by the system. Consequently, AI systems may execute malicious instructions, posing significant threats to personal security and privacy.

Autonomous Driving (AD) systems are another well-representative type of AI system, which also exerted great interest in recent years. However, these systems face multiple system-level vulnerabilities, ranging from hardware to communications. For instance, Autonomous Vehicles (AVs) utilize a variety of sensors, such as cameras and LiDAR, to gather environmental information. These sensors act as the "eyes" of the vehicle. However, the integration of new sensors like LiDAR can introduce hardware-level vulnerabilities – Cao et al. (2019) highlighted how attackers could exploit these vulnerabilities by using laser equipment to create deceptive "virtual" point clouds, potentially compromising Vehicle Automation (VA) systems. Beyond sensor devices, Hong et al. (2020) demonstrated the feasibility of executing malicious programs in tandem with AI components in the Robot Operating System (ROS). Building upon this, Tu et al. (2021) have shown that the communication protocols in multi-agent VA systems can be attacked through adversarial shared information via a black-box transfer attack.

In AI security, *traditional* cyber-security threats also come into play. A malicious actor could obtain the weights of a model via non-AI-related techniques, e.g., by stealing the credentials of a company employee via social engineering techniques. This could have significant consequences as it may simplify the execution of several of the attacks presented above (for instance, evasion attacks). Moreover, the software used to train and deploy AI should be treated as carefully as all other software: for example, loading untrusted PyTorch models via the widely used `torch.load` function can be dangerous as it allows for arbitrary code execution[11].

C Other Relevant Disciplines

While this paper predominantly focuses on AI safety and AI security, we also acknowledge the importance of multiple other disciplines that are closely relevant to AI risk management. Underlying these disciplines are a diverse set of communities with rich literature, however, an exhaustive survey of all these topics is beyond the scope of this paper. This section supplements a brief discussion on a few such topics that are particularly salient in the existing literature, meanwhile, we also acknowledge other related topics not covered.

Fairness generally concerns the equitable and just distribution of both the benefits and risks of AI technologies. In practice, standards and definitions of fairness can greatly vary across different applications (Barocas et al., 2023). The definitions of fairness can also be subjective (Grgic-Hlaca et al., 2018), culturally influenced (Greenberg, 2001), and even conflicting (Kleinberg et al., 2016). The intersections of AI fairness with safety and security considerations are noteworthy. Deploying LLMs without proper consideration of fairness could perpetuate existing biases and stereotypes prevalent on the internet and in society (Welbl et al., 2021), a concern that is often in safety considerations. Regarding security, adversaries can easily manipulate the system's behavior to introduce bias (Solans et al., 2020; Chang et al., 2020; Mehrabi et al., 2021). More broadly, the fairness implications of AI on the external environment can be highly complex and multifaceted, often *extending beyond a simple binary understanding of being solely benign or harmful.* Numerous AI applications may involve multiple stakeholders with diverse and sometimes competing interests. For instance, companies can reduce expenses in hiring processes using AI-driven job recruitment, and financial institutions may benefit from streamlined evaluation processes such as credit scoring algorithms — even if these AI techniques are not inherently fair. For the objects being evaluated by these AI algorithms, individuals who fit particular data patterns can directly benefit from "an unfair AI" though disadvantaging other minority groups with non-traditional backgrounds (Dastin, 2022; Fuster et al., 2017).

Privacy also presents a rather complex challenge in AI risk management. It is a comprehensive concept that plays a crucial role in protecting fundamental values like human autonomy and dignity. Although the definition of privacy (Solove, 2010) and strategies for its protection vary across cultural contexts, individual preferences, and application scenarios, several common privacy-enhancing techniques are widely recognized. These include, but are not limited to, anonymization (Bayardo & Agrawal, 2005), removing personally identifiable information (Administration, 2019), differential privacy (Dwork et al., 2014), and secure multi-party computation (Yao, 1986; Goldreich, 1998). Privacy-related risks encompass a broad spectrum that intersects with safety and security concerns to varying degrees. Certain privacy breaches directly expose safety or security risks, like chatbots accidentally disclosing user data (jutogashi, 2023; Cole, 2023) and intentional data extraction attacks by adversaries (Carlini et al., 2021, 2023a; Nasr et al., 2023a). On the other hand, some privacy risks such as membership inference (Shokri et al.,

[11] https://github.com/pytorch/pytorch/issues/52596

2016) and the auditing of certain privacy mechanisms (Jagielski et al., 2020b; Steinke et al., 2023) primarily pertain to traditional privacy concerns, but they can also indirectly cause safety and security vulnerabilities (Yao et al., 2023). A thorough exploration of privacy requires a more nuanced discussion, which falls beyond the scope of this paper. For an in-depth understanding of privacy definitions and risks, we recommend referring to the NIST Privacy Framework (NIST, 2020) and a recent survey by Cummings et al. (2024).

Copyright. Another emerging concern in AI risk management is the issue of copyright. Generative AI's role in content creation offers significant benefits in creativity and productivity, yet poses risks to copyright holders – The advancement of generative AI could disrupt traditional markets, where original content creators might struggle to compete with AI-generated content that leverages their work without appropriate compensation or acknowledgment (Henderson et al., 2023b; Deng & Ma, 2023).