The Pitfalls of "Security by Obscurity" and What They Mean for Transparent AI

Peter Hall[1], **Olivia Mundahl**[1], **Sunoo Park**[1]

[1]New York University
{pf2184, om2145, sunoo.park}@nyu.edu

Abstract

Calls for transparency in AI systems are growing in number and urgency from diverse stakeholders ranging from regulators to researchers to users (with a comparative absence of companies developing AI). Notions of transparency for AI abound, each addressing distinct interests and concerns.

In computer security, transparency is likewise regarded as a key concept. The security community has for decades pushed back against so-called *security by obscurity*—the idea that hiding how a system works protects it from attack—against significant pressure from industry and other stakeholders, e.g., (Bellovin and Bush 2002). And over those decades, in a community process that is imperfect and ongoing, security researchers and practitioners have gradually built up some norms and practices around how to balance transparency interests with possible negative side effects. This paper asks: *What insights can the AI community take from the security community's experience with transparency?*

We identify three key themes in the security community's perspective on the *benefits of transparency* and their approach to balancing transparency against countervailing interests. For each, we investigate parallels and insights relevant to transparency in AI. We then provide a case study discussion on how transparency has shaped the research subfield of anonymization. Finally, shifting our focus from similarities to differences, we highlight key transparency issues where modern AI systems present challenges different from other kinds of security-critical systems, raising interesting open questions for the security and AI communities alike.

Introduction

Artificial intelligence has proven to be highly impactful, with impacts in critical domains such as biosciences (Jumper et al. 2021), health (Rajpurkar et al. 2022), and public safety (Fine and Marsh 2024). It is also continually reaching new peaks in consumer and commercial interest, with foundation models such as Claude, Llama, Stable Diffusion, and the GPT family posited to be adaptable to diverse uses (Jain, Maleki, and Saade 2024). Despite their already widespread impacts, there is still much to understand about how recent developments in AI and machine learning models operate, and the broader human consequences their use and misuse may have. In the commercial sphere, companies have generally (though not always) preferred to hide their methods of data collection, training, fine-tuning, and other specifications (Carlini et al. 2023).

At the confluence of these circumstances, many are invested in achieving (various notions of) *AI transparency*. Researchers want to investigate what makes AI models, including production-line models, better or worse at various tasks (e.g., (Bender et al. 2021)), and to understand unexpected or possibly harmful effects of model use (e.g., (Awasthi et al. 2023; Carlini et al. 2019; Raimondo and Locascio 2023)). Consumers, businesses, and other clients should have assurances that companies selling AI-based tools are able to achieve what they are promising—and even before that, they should be able to understand what they are being promised. As increasingly consequential decisions are made with contributions from AI models, people impacted by these decisions need assurances about how AI impacts their lives (e.g., in healthcare (Khan et al. 2023) or bail setting (Fine and Marsh 2024)). Copyright holders want notice and fair compensation and attribution when others are using their works–often making large profits (S.D.N.Y. 2023; Gilbertson and Reisner 2024). Regulators want to understand and protect their constituencies from potential harms (Weidinger et al. 2021) while promoting innovation and competition in AI (in the Office of Technology 2024a,b; Vestager 2024). Meanwhile, skeptics have variously argued that too much transparency will harm innovation, business interests (including trade secrets), data privacy, system security, national security/defense, and/or public safety (e.g., (Patel and Toomey 2024; Hosanagar 2024)).

An important emerging literature in both research and policy has responded to this ongoing debate by investigating transparency in AI. Some have been trying to build in properties like explainability or interpretability to LLMs (Graziani et al. 2022). Others are hoping to balance data privacy with open source models (in the Office of Technology 2024a; Zuckerberg 2024). What this may mean is the subject of ongoing debate (Gibney 2024). Others take a legal or policy approach, like the Biden Administration's Executive Order on Trustworthy AI and the European Union's AI Act (Biden 2023; Union 2024).

The idea that *hiding how a system works protects it in some way* is a key recurring theme in the current dis-

course on transparency in AI. This same idea is, however, one whose promises and pitfalls the security community is deeply familiar with (Kahn 1996). Attempts to protect systems by hiding how they work have prompted the security community's participation in heated debates across academia, industry, and policy over decades (e.g., (Bellovin and Bush 2002; Diffie 2003; Mercuri and Neumann 2003; Schneier 2004; Shipman 2019)). We seek to understand the themes across the debates unfolding in AI and those longer entrenched in security.

Our paper thus offers a novel and systematic exposition of the essential principles underpinning transparency in security, which we believe have useful parallels to transparency in AI. That is, the central question of this paper is:

What insights can the AI community take from the security community's experience with transparency?

Transparency in Security

In modern computer security, a hard-fought broad-based consensus has been established: *Despite the intuitive idea that hiding a system should protect it, transparency is often more beneficial for protection*. The consensus on this general principle is broad, though perspectives on how to implement the principle in specific contexts can be more varied.

It was not always this way. Before modern cryptography and systems security, even critical military communications often relied on *security by obscurity* (e.g., (Tohe 2022; Katz and Lindell 2015)): namely, the idea, now well established to be fallacious, that hiding how a system works is an effective and adequate defense against adversaries.

Moving away from security by obscurity has been a long and ongoing process.[1] Now, "security by obscurity" is a catchphrase in the security community with negative connotations, and relying upon security by obscurity is widely considered inadequate (Bellovin and Bush 2002). Perfect unanimous agreement on the precise type of transparency that is best for every type of system and application context may never be reached—yet the received wisdom that transparency promotes better security practices and quicker mitigations of harm has held strong, and served as a source of unity in the community.

We call this modern approach *security by transparency*—a natural phrase at times used colloquially to describe the opposite of security by obscurity. As the security-by-transparency approach has developed through community experience and norm-building over time, a substantial part of it lies in folklore and institutional memory, written records of which are relatively sparse and scattered.

Security by Transparency

The notion of security by transparency predates computer security. As far back as the 19th century, Auguste Kerckhoffs (Kerckhoffs 1883), writing about how to design secure military communication, remarked that a truly secure system should remain secure even assuming the enemy knows everything about a system. That is, systems should *by design* be robust even when the system design is known.

The argument for security by transparency is subtle: The idea that hiding how a system works can help protect it is not only intuitive, it is in a technical sense *correct*. If you build all the best state-of-the-art security measures into your system and then choose not to disclose how it works at the end, it may in fact be harder for attackers to find vulnerabilities in the system, as they'll first have to figure out how it works. (And if you *could* build a perfectly secure system, whether or not you disclose how it works doesn't matter; the security will be perfect either way. But we don't know how to build perfect systems; and the security community has embraced this reality as a critical engineering consideration (Viega and McGraw 2001).)

Why, then, is security by obscurity so poorly regarded?

First, *relying* on obscurity for security is risky at best, and creates a false sense of security at worst. The phrase "security by obscurity" sometimes refers specifically to such *reliance*: Having obscurity as the main protection, rather than as one measure among many. Kerckhoffs' Principle warns directly against this. Second, decades of experience have shown that opacity leads to externalities that tend to indirectly undermine the robustness of designs and systems.

We stress that security by transparency is a paradigm, not a rigidly defined set of rules. In a given situation, security by transparency provides principles with which to reason about what to disclose to whom, what to keep secret, and the possible consequences. These principles do *not* prescribe disclosure of every last detail of a system, or prescribe any singular foolproof way to achieve security.

Our Scope: Transparency, Not Security

Our focus is not on *security* for AI systems but rather on *transparency* for AI systems. We are interested in the parallels and potential for cross-pollination between the security and AI communities in their debates around transparency.

When we talk about security by transparency, our discussion will often center around how principles of transparency promote *security* in some way—naturally, as the context is the field of security. We argue, however, that these discussions' relevance to the AI community extends beyond securing AI, to AI transparency much more broadly—for at least two reasons. First, we are focused on high-level principles, and security at a high level is simply about "building systems to remain dependable in the face of malice, error, or mischance" (Anderson 2021)—the very same concerns underlying many of the calls for transparency in AI, including some we might not usually think of as "security concerns." Second, transparency aids security *as a discipline* in ways that may translate to other disciplines: For example, allowing researchers to uncover more efficient solutions and richer properties and functionalities of systems, and to better understand system guarantees and limitations.

Background

Beginnings of Transparency in Security: By the late 19th century, cryptographers had reservations about security

[1] Obscurity is still common in some contexts, e.g., Digital Rights Management (Liu, Safavi-Naini, and Sheppard 2003).

by obscurity. In the seminal 1883 work "La Cryptographie Militaire" (Kerckhoffs 1883), Kerckhoffs voiced six principles he believed essential to strong security systems. Among them, known now as Kerckhoffs' Principle, stated approximately: "The system must not require secrecy and can be stolen by the enemy without causing trouble."[2] Thus began the counter viewpoint, security by transparency.

Disclosure Practices: From the earliest days of consumer computer systems, researchers and concerned citizens have been finding and sharing found vulnerabilities. In the late 1980s and '90s, this was through email chains and online zines like Bugtraq and Security Digest[3]; now, security experts are more likely to communicate vulnerabilities to the developers through formal bug bounty processes and to the public through peer-reviewed publications.

There has been much debate over the best way to disclose vulnerabilities both to developers and the public. The details of this debate are beyond our scope;[4] here, we just summarize a few key terms related to vulnerability disclosure.

Full disclosure involves sharing all information about an attack with the public immediately. Many are concerned that full disclosure results in too much potential for harmful exploitation of attacks (Culp 2001). *Coordinated vulnerability disclosure* (CVD) involves an initial private disclosure followed by publication after a delay (CERT 2024).[5] Increasingly, many organizations have groups who handle bug and vulnerability reporting,[6]; this streamlines CVD processes and timely fixes. *Bug bounty programs*, where community members are invited to submit found vulnerabilities to the company for money or perks, are a way for organizations to proactively seek disclosures.[7]

Attacks and Vulnerability Papers: At security venues, attack papers are now common, though they were more controversial in the past. Basin and Capkun (Basin and Capkun 2012) provide one account of the case for attack papers: Finding and fixing bugs, learning flaws which may guide research, understanding compatibility or lack thereof, finding exact security guarantees.

Calls & Goals for Transparent AI: Calls for transparency in AI are numerous (e.g., (White et al. 2024)). Notable relevant terms and areas of study include: explainability, interpretability, accountability, trustworthiness, and robustness (NIST 2023). Recent advances are making important progress in AI modeling (Dev et al. 2024) but often focus on defining largely positive guarantees. Additionally, some risks can be difficult or elusive to define in a theoretically sound way (Polemi et al. 2024). Disclosure practices are being developed (Argentieri 2024; Cattell, Ghosh, and Kaffee 2024), but researchers remain at risk in AI due to uncertainties and discrepancies regarding policies (Tiku 2024). Prominent motivations for these efforts towards transparency include public trust (Andrada, Clowes, and Smart 2022; Schneier 2023), technological advancement (Fukawa, Zhang, and Erevelles 2021; Bostrom 2018), mitigation of bias (Carlson 2017; Ferrara 2023), and liability (Novelli, Taddeo, and Floridi 2023; Cooper et al. 2022).

We refer to the full version for a discussion of black-box access, post-hoc and ante-hoc transparency, and open source.

Connections In Transparency: From Security To AI

This section highlights three key themes underpinning security by transparency, with discussion of parallels with AI.

Modeling is Essential For a Robust Understanding of Systems That Are Too Complex To Fully Model

Perhaps paradoxically, modeling serves as an essential tool to maintain a robust understanding of what we can *and cannot* guarantee about the performance and failure modes of systems that are too complex to fully model — which have (security) requirements too complex to fully model.

But wouldn't modeling such a system (and its requirements) be futile as any model would fail to capture important aspects of real-world operation? Even worse, wouldn't modeling such systems rather lead to dangerous oversimplifications and misunderstandings? These are reasonable questions: oversimplified models do suffer from these serious drawbacks *when their limitations are not understood*. In contrast, the type of modeling for which we advocate critically *includes* clear characterization of the model's own limitations: that is, modeling what we can *and cannot* guarantee about a system on the basis of a given model.

Security engineering involves building systems with stated *threat models* and *assumptions*, which delineate (1) what functionality, confidentiality, and robustness guarantees a system is designed to have (2) against which kinds of threats (3) under what conditions and, implicitly or explicitly, (4) which kinds of functionality are *not* guaranteed and what kinds of threats are *not* protected against. Threat modeling has proven an essential technique to build (necessarily) imperfect systems that have clear specifications of *what is guaranteed* and *what is not*, and in what scenarios these guarantees are assured. The threat model is generally considered part of the system specification: it is seen as necessary to understand "what's in the box," as well as how (not) to use it and for which application contexts it is appropriate.

Transparency plays a key role here in several ways. First, if an assumption necessary for a system's security is proven incorrect, clearly stated assumptions allow for the community to quickly adapt by pruning ideas which rely on it. Second, by making these assumptions explicit, we encourage research to find ways to test or bolster the assumption,[8] as well as make it more likely that the research community will

[2] In the original French: "Il faut qu'il n'exige pas le secret, et qu'il puisse sans inconvénient tomber entre les mains de l'ennemi."

[3] https://seclists.org/bugtraq and securitydigest.org, respectively

[4] We refer to plenty of literature on the topic (e.g., (Moura and Heidemann 2023; Wicker 2021; Moura and Heidemann 2023)).

[5] The exact details of appropriate CVD timeline are context-dependent and a subject of ongoing community discussion.

[6] See, e.g., the Microsoft Security Vulnerability Research group.

[7] See, e.g., Bugcrowd.

[8] E.g., the fallibility of passwords in practice gives rise to other authentication schemes like biometrics, 2FA, etc.

timely find out if an assumption does not hold. Third, it enables both developers and users to plan ahead for system failures in a way tailored to the threat model. Fourth, it aids stakeholders to evaluate in what contexts the use of a tool is riskier, and in what contexts not to use a tool at all. Finally, clear assumptions and threat models serve as stepping stones for further research and more refined models that further enhance the above benefits.

All of the above points have baked into them the implicit idea that we cannot "simply" build a system to be secure, not only because we do not know how to build perfect systems but because we cannot write a single specification of "security." Security objectives are context-dependent, and full specification is often intractable. In sum: (1) there is no general-purpose certification of "secure enough;" (2) we aim to design threat models appropriate to limited contexts, stress they are *not* general purpose, and clearly define what they do *not* cover; and (3) we associate clearly stating and disseminating such threat models with the benefits above.

Connection to AI Some simple AI tasks are relatively clear and self-contained in their goals. But in many settings, the goals can often be less clear, expressed as heuristics, inherently incomplete, inherently sociotechnical (and thus not amenable to complete technical specification), and/or context-dependent. Explicit "threat" modeling in the AI context can embrace this incompleteness and the idea that *these systems and requirements are too complex to fully model*, and seek to delineate what is not clear, what is incomplete, and what can(not) be guaranteed. By doing so, it has the potential to play a similarly critical role in reinforcing robustness for AI systems deployed in critical contexts when robustness is a context-dependent moving target.

None of this is to say that AI research is devoid of modeling — far from it. Recent positive strides include theoretical modeling (e.g., (Górriz et al. 2023)), candidate definitions of robustness, fairness, and privacy (e.g., (NIST 2023)), transparency in experimental design (e.g., (Felzmann et al. 2020)), and transparency of societal impacts (e.g., (NIST)). These efforts make important progress on defining *positive guarantees* but we have seen less depth of engagement with *negative guarantees*: Precisely characterizing the types of situations and failure modes *not* addressed by proposed models and definitions. Here, we mean going beyond the important initial step of mentioning that certain enumerated things out of scope of one's model; we mean, rather, striving towards a *complete* characterization of what's not in scope — such that if anyone thinks up a threat, they can straightforwardly ascertain whether and how it is in scope, by reading a concise threat model. We recognize that this is a complex goal that will not be achieved overnight, and that important research has already been done in this direction (e.g., (Stadler et al. 2024)); our aim is not to criticize but to highlight this natural and fruitful avenue for future progress.

"Many Eyes" & Disclosure

A key impact of transparent practices within security has been the ability to more efficiently and reliably correct errors, vulnerabilities, and other problems. Examples of such transparent practices include bug bounty programs, cryptographic standardization processes, and the publication and recognition of attack papers. These practices all invite scrutiny from all over, letting the entire community contribute. Underpinning this reasoning is what is known as the "Many Eyes" Theory, also known as Linus's Law: "Given enough eyeballs, all bugs are shallow" (Raymond 1999). That is, giving all stakeholders the ability and incentive to investigate a system aids in better design and mitigation.

Positively receiving and proactively seeking community input on security issues are essential practices that have become much better established over time, although these were rare practices in earlier days (and are far from ubiquitous even today; progress is ongoing). RSA (Administrator 1991) famously prompted the community to find weaknesses in their systems and share that information for prizes (eventually claimed, like in (Cavallar et al. 2000)). Now, many companies solicit vulnerability reports or offer "bug bounties" for them. Organizations that set standards and offer community and support for soliciting reports are also influential and growing. [9] These recent trends (though not perfect) starkly contrast with early examples of hostile reactions to reports of security issues (e.g., (Foundation 2008, 2013)).

Realizing the full potential benefit of community input is challenging in practice as it requires trust and mutual understanding between stakeholders with incentives that are not always aligned: the community providing input and the community receiving input (though these are not always disjoint). Lack of established precedents and collaborative norms can lead to uncertainty and tensions, as both the security and AI communities have experienced—for security, more acutely in earlier decades. These, in turn, impede constructive progress based community inputs. The security community has, over time, developed stronger collaborative norms around disclosure processes that facilitate mutual benefits from community inputs; that said, the process is imperfect and norm-building is very much ongoing.

When some flaw in a system is suspected and found, choices must be made about who to tell when (if at all). While specific standards vary, it is a widely held belief within the security community that there should generally be some form of a graded disclosure process rather than a single public disclosure, but that public disclosure should eventually happen and generally promotes security.

When a report of a flaw is received, choices must likewise be made about how to respond and how to address the report (if at all). Only when the one side's approach to who to tell when and the other side's approach to how to respond and address the report are aligned can the full potential benefit of community input be realized. And yet, incentives are not always aligned between the two sides.

Increasing community norms around disclosure processes have made progress toward aligning the two sides' approaches. The community's acceptance of and collaboration on vulnerability reports today stands in notable contrast to earlier decades, where security researchers were more often subject to aggressive legal and reputational attacks in reac-

[9] E.g., disclose.io, Hackerone, and Bugcrowd.

tion to their findings (e.g., (Rauch 2022; Foundation 2008; Gamero-Garrido et al. 2017)). The difference between research that is beneficial for security and *maliciously attacking* a system—confusion over which was a recurrent feature earlier on—became better understood and more clearly distinguished over time, by stakeholders in both industry and research (e.g., (Rauch 2022)). Avenues for coordinating disclosure continue to expand, e.g., involving a government intermediary (Specter, Koppel, and Weitzner 2020, §6).

That said, much progress remains to be made. Terms of service often still prohibit security researchers from accessing and analyzing systems for research purposes. Security researchers continue to face legal and reputational attacks over their research, and navigating disclosure processes can still involve tensions and risk (Park and Albert 2024; Moura and Heidemann 2023; Specter, Koppel, and Weitzner 2020).

Connection to AI At present, there are many barriers to implementing similar processes in AI. The terms of service of many prominent AI tools disallow or disincentivize the work needed to study vulnerabilities and the disclosure of such flaws (Longpre et al. 2024). AI researchers have been targeted for their research by developers; some have been instructed to redact certain parts of their research or findings (Carlini et al. 2024). These barriers feel very familiar to the experiences and history of the security community. And in security, it is these early barriers that led to the development of today's disclosure processes.

To ensure that AI systems are evaluated and properly understood, the security community's experience suggests that promoting an active research community and access by the research community can in fact promote a more robust end product with better understood guarantees.

Also, a clear, codified disclosure process could go a long way to protecting any private information while also increasing trust between the community and developers. Some have attempted to explore what this could look like (Rando et al. 2022; Longpre et al. 2024), and we believe this is critical to transparency for AI. Of course, as the security community's experience also illustrates, building community around such processes takes time.

Ubiquitous Technologies Necessitate Public Trust

A common concern about transparency (in security and AI) is that average users cannot generally be expected to understand the technologies involved even if they are carefully explained in full detail—so then, what purpose does the transparency serve? Laudable efforts have been made to promote understandability for lay users, but the full complexity—and the range of risks and failure modes—of these technologies will not realistically be understood by every user impacted.

Despite the fact that the average person is not expected to understand security best practices or how encryption works, transparent practices in security benefit everyone. Whether they are aware of it or not, everyone who uses computers uses security features and engages in compliance with security best practices—they are built in at a hardware level, at a protocol level, and at a networking level. This is facilitated by transparent practices allowing for users and developers to buy in to best practices. The security community's experience shows that the ubiquity and broad impact of systems where security matters necessitates and supports the widespread adoption of public trust mechanisms, *even when* user understanding and market demand for transparency appears lacking, and companies seem to lack incentives to be transparent. The result is more reliable and coordinated security practices across complex interconnected systems.

Many commonplace interactions online require users to rely upon the safety and robustness of systems where security matters. By using these systems, users place implicit trust in the systems; and the more essential the use of the system is to modern life, the less choice they have about placing their implicit trust in the systems. And sometimes, the choice to use the system in a way that impacts a given person is made by someone else (e.g., an employer).

Thus, people implicitly trust that their credit card information will not be stolen when they enter it in an online form. They trust that their messages are private and go to the intended recipient. They trust that their webcams are not on when they're not using them, and they trust that their smart watch monitoring their health condition will flag anomalies. They trust that their employer's payroll system works. Such trust does not come directly from individuals studying transparent literature to understand design practices, but rather, indirectly through public trust mechanisms.

By public trust mechanisms, we refer to a host of approaches. For example, in fields like civil engineering, certifications and regulation serve as public trust mechanisms. In biomedical sciences, extensive testing is another public trust mechanism. Fields like psychology and physical sciences have academic standards and checks as public trust mechanisms. In these cases, the stakes are clear: People could die if things are not done safely.

When considering the safety and robustness of complex computing systems, though, the potential harms often seem more intangible. This lack of visibility can lead to a lack of awareness among lay users as well as a lack of incentive for companies to improve either security or transparency—in economic terms, creating a *moral hazard* (Vagle 2020).

We see strong transparent practices as a way that security researchers have built up community-wide public trust mechanisms, that in turn promote systemic deployed security that benefits everyone notwithstanding all of the countervailing forces and considerations discussed above.

Connection to AI The ongoing debate on AI transparency is likewise grappling with the facts that the average user cannot be expected to understand AI or fully evaluate the benefits and risks associated with its use, that companies generally seem to lack incentives to be transparent, and that users do not seem to create market demand for transparency.

AI is also growing rapidly, being deployed in a wide range of contexts with broad impacts, and is poised to be embedded into technologies at a scale that leaves users little choice around whether to engage with it. In many cases (including, e.g., policing (Angwin et al. 2016) and welfare distribution (Zouridis and Bovens 2019)), people cannot opt out of algorithmic decision-making, including in cases where these

may have serious consequences.

Indeed, many have taken notice of the prevalence of AI in their lives, and many have raised concerns. However, the incentives are not necessarily aligned for any one group to contend with these issues, even if resolutions are in the public's interest. The combination of this kind of *ubiquity and impact* with the *lack of understandability and tangibility to users*, and related *weak incentives for companies* creates the conditions—shared between AI and security—that we believe necessitates transparent practices within community-wide public trust mechanisms. These, in turn, promote systemic safety and robustness measures that can benefit society despite the complex incentives involved.

We believe transparency is necessary as a public trust mechanism for secure systems and robust AI alike, both for the benefit of the public and also for the adoption and continued use of these technologies to their full potential.

A Case Study on (De)-Anonymization

We believe the security-by-transparency mindset has shaped how each research area within the broad umbrella of security has developed. Here, we give a brief summary of one such domain which may be instructive: Anonymization.

At a high level, anonymization refers to methods designed to hide individually identifiable information in a dataset while retaining useful data accesses and statistics. For some time, anonymization was pitched as the answer to being able to make use of the vast datasets of private information that online platforms and governments amass (e.g., communications, content consumption, or census data), while also protecting individual safety and privacy (Ohm 2009; Rubinstein and Hartzog 2016). Proposed anonymization techniques were taken by some as a green light to "anonymize" sensitive datasets and then use them for any purpose, including publishing them. A series of works around the 2000s exposed this as overly optimistic (e.g., (Barbaro and Jr. 2006; Narayanan and Shmatikov 2008; Sweeney 2000)), de-anonymizing willingly published user data from sources such as AOL and Netflix. The demonstration of viable attacks was what prompted stakeholders to take the risks more seriously, and catalyzed change in community practices in industry, research, and beyond.

On research methodologies and norms, we learned how to investigate compromised datasets ethically (Bonneau 2012) while taking into account privacy and anonymity considerations, and studied whether and how to use datasets of illicit origin for research (Thomas et al. 2017).

On modeling, Dwork et al. (Dwork 2006; Dwork et al. 2006) introduced the notion of *differential privacy* (DP), a notion which has also featured impactfully in machine learning research (e.g., (Yu et al. 2024; Xian et al. 2024)). Dwork discusses DP from a theoretical standpoint, including explanations for modeling decisions, and the DP framework enables precise mathematical reasoning and tradeoff-making about certain kinds of deanonymization risks.

On awareness and adaptation, deanonymization research since the late 1990s gradually impressed upon computer scientists and the broader public that anonymization techniques are unreliable and poorly understood—and later on, that not a single one provided perfect anonymity under researchers' scrutiny, and anonymization might be a pipe dream (Ohm 2009). Organizations now appear less likely to optimistically publish datasets that would be damaging if deanonymized. Subsequent works explore whether and how anonymization techniques could be useful despite acknowledged imperfections, if tailored to specific applications (Angiuli, Blitzstein, and Waldo 2015).

Through transparent research about anonymization's flaws, the community has been able to embrace the idea that this goal is infeasible, and to develop reasoned methods of handling datasets and privacy risks accordingly.

Novel Challenges

This section highlights significant differences between security and AI that may necessitate new approaches. We refer to the full version for more discussion on novel and perceived challenges[10].

Training Data: In security, there is usually a clear delineation between (1) a system's design and functionality and (2) (private) input data within the system. In AI, this line is blurred, with the training data for a model being inextricably linked to its performance and properties. Just the untrained model is insufficient for analyzing its properties (Felzmann et al. 2020). Additionally, some key motivations for transparency, such as model bias (Suresh and Guttag 2021) and memorization (Carlini et al. 2023), contend explicitly with training data and its expressions in the trained model.

Some AI models could reasonably make their training data public. However, disclosing the training set may expose developers and researchers to serious legal and ethical risks. To mitigate these concerns, some have proposed disclosing only (partial) model weights, though this still carries risk of exposing training data (Nasr, Shokri, and Houmansadr 2019; Nasr et al. 2023). Some have proposed anonymized or synthetic training data; these approaches carry their own risks (Narayanan and Shmatikov 2008; Cristofaro 2024).

Security has not had to face the same issue. We believe this is an important open problem to solve, and one in which collaboration between security and AI researchers may be a fruitful avenue to understanding training-data confidentiality concerns and how they interact with transparency (Carlini et al. 2023; Nasr et al. 2023).

Disclosure Processes: It is important to consider how the disclosure processes may be adapted for the AI context. We caution against simply adopting the exact same procedures as security. The entanglement of training data with system design makes the relevant stakeholder set in AI larger and potentially unclear. This is especially complicated as the interests of stakeholders may be in tension.

While important initial progress has been made on frameworks for AI vulnerability disclosure (Raimondo and Locascio 2023; Cattell, Ghosh, and Kaffee 2024), the risks and at-risk parties implicated by different vulnerabilities vary wildly. Understanding and careful modelling of what is and

[10]https://arxiv.org/abs/2501.18669.

is not inherent to AI and machine learning techniques would help piece through these, but it is possible the scope of vulnerabilities and thus responsible disclosure in AI may be an inherently more complex problem.

Brittleness of Models: The well-documented and sometimes inherent tradeoffs between privacy (of training data) and performance is a key feature of AI development (Carvalho et al. 2023). In security, there is almost always a tradeoff between efficiency and privacy, and this is the main tradeoff to consider.[11] Choosing the right balance of privacy and performance is an inherently non-technical problem (Schneier 2003), and sometimes, development toward privacy and performance fundamentally contradict each other (Gu et al. 2022). The opportunity for technical work is in the modeling of the tradeoff. While there are parallels between the two communities, it seems the tradeoff problem in the AI community is certainly different in scale and possibly different in kind (e.g., involving multidimensional optimization (Monteiro and Reynoso-Meza 2023)), posing a greater challenge in finding an appropriate tradeoff—and heightening the possibility of accidental domino effects.

Optimizing for Metrics: Many AI systems are based on optimizing for metrics serving as heuristics for a real-world objective. Metrics feature across development, from initial training to the addition of post-hoc "guardrails." Disclosing metrics can make them easier to game (Goodhart 1975). These observations would seem to counsel against wide disclosure of metrics (such as objective functions) in order to preserve the utility of the metrics and thus the utility of the system. Yet metrics, like training data, constitute an essential part of system functionality. Analyzing a system with the metrics redacted is likely to result in a significantly incomplete understanding of its functionality. And some key motivations for transparency explicitly contend with the metrics used. The idea that releasing the full details of system functionality can inherently make the system *less useful* for its intended purpose does not have a good analogue in security.

Perceived Challenges

We briefly mention arguments against transparency in AI which have parallels to those in security. We refer to the full version for more complete discussion and more references.

Trade Secrets and Innovation: One common concern is that too much disclosure of algorithms and data would reduce incentives for companies to innovate through costly investments in AI (e.g., (Hind et al. 2020; et al. 2022))—thus harming quality of AI overall as well as competitiveness in a global market. Likewise, industry stakeholders have long used intellectual property and innovation as reasons to argue against transparency of technology in security contexts (e.g., (Chakraborty and Bhunia 2009)).

Misuse: Another common concern is that releasing AI systems may lead to bad consequences from their deliberate or accidental use for harm, and that the potential or likelihood of such harm may increase with fuller disclosures (e.g., (Gade et al. 2024)). These narratives have limited parallels to long-standing discussions in the security community around privacy tools (such as encrypted messaging, Tor, and cryptocurrencies) being misused.

Pace of Development: Many have commented on the unusually fast pace of recent AI developments (e.g., (Bengio et al. 2024)). This pace has naturally impeded the establishment of standards and best practices (Bostrom 2017). While the security community has not experienced a comparable pace of development, it has still slowed down compared to an earlier phase, based on a collective experience of flaws discovered in systems that were developed too quickly and scrutinized insufficiently before deployment.

Consequential Outcomes and a Right to Know: As AI systems are used in ways that have increasingly consequential outcomes, some calls for transparency center the idea of a "right to know" about decisions that impact individuals or groups (e.g., (Fehr et al. 2024)). Similar discussions have arisen in the context of security-critical systems that impact consequential outcomes (e.g., in election security and government technology (Girgvliani 2023; Jones 2007)).

[11] There are contexts in security with more nuanced tradeoffs, e.g., encrypted search (Curtmola et al. 2006; Boneh et al. 2015) and differential privacy (Dwork et al. 2006; Dwork and Roth 2014).

Acknowledgments

We are grateful to Kyunghyun Cho, Ana-Maria Cretu, Andrés Fábrega, Betty Li Hou, Robert Mahari, Falaah Arif Khan, Adam Sealfon, Vitaly Shmatikov, and our anonymous reviewers at AAAI 2025.

References

Administrator, R. C. 1991. RSA Factoring Challenge Announcement.

Anderson, R. 2021. *Security Engineering*. Wiley, 3rd ed. edition.

Andrada, G.; Clowes, R. W.; and Smart, P. R. 2022. Varieties of transparency: exploring agency within AI systems. *AI and SOCIETY*, 38: 1321–1331.

Angiuli, O.; Blitzstein, J.; and Waldo, J. 2015. How to De-Identify your Data. *Communications of the ACM*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. ProPublica.

Argentieri, N. M. 2024. Principal Deputy Assistant Attorney General Nicole M. Argentieri Delivers Remarks at the Computer Crime and Intellectual Property Section's Symposium on Artificial Intelligence in the Justice Dept. U.S. Dept. of Justice.

Awasthi, P.; Mao, A.; Mohri, M.; and Zhong, Y. 2023. Theoretically Grounded Loss Functions and Algorithms for Adversarial Robustness. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 10077–10094. PMLR.

Barbaro, M.; and Jr., T. Z. 2006. A Face Is Exposed for AOL Searcher No. 4417749.

Basin, D.; and Capkun, S. 2012. The Research Value of Publishing Attacks. *Communications of the ACM*, 55: 22–24.

Bellovin, S. M.; and Bush, R. 2002. Security Through Obscurity Considered Dangerous. The Internet Society.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. ACM.

Bengio, Y.; Hinton, G.; Yao, A.; Song, D.; Abbeel, P.; Darrell, T.; Harari, Y. N.; Zhang, Y.-Q.; Xue, L.; Shalev-Shwartz, S.; Hadfield, G.; Clune, J.; Maharaj, T.; Hutter, F.; Baydin, A. G.; McIlraith, S.; Gao, Q.; Acharya, A.; Krueger, D.; Dragan, A.; Torr, P.; Russell, S.; Kahneman, D.; Brauner, J.; and Mindermann, S. 2024. Managing extreme AI risks amid rapid progress. *Science*, 384.

Biden, J. R. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. EO14110.

Boneh, D.; Lewi, K.; Raykova, M.; Sahai, A.; Zhandry, M.; and Zimmerman, J. 2015. Semantically secure order-revealing encryption: Multi-input functional encryption without obfuscation. In *EUROCRYPT*, 563–594. EUROCRYPT.

Bonneau, J. 2012. The Science of Guessing: Analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy*. IEEE Symposium on Security and Privacy.

Bostrom, N. 2017. Strategic Implications of Openness in AI Development. *Global Policy*.

Bostrom, N. 2018. Strategic Implications of Openness in AI Development. *Artificial Intelligence Safety and Security*.

Carlini, N.; Athalye, A.; Papernot, N.; Brendel, W.; Rauber, J.; Tsipras, D.; Goodfellow, I.; Madry, A.; and Kurakin, A. 2019. On Evaluating Adversarial Robustness. *arXiv*.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramèr, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting Training Data from Diffusion Models. *arXiv*.

Carlini, N.; Paleka, D.; Dvijotham, K. D.; Steinke, T.; Hayase, J.; Cooper, A. F.; Lee, K.; Jagielski, M.; Nasr, M.; Conmy, A.; Yona, I.; Wallace, E.; Rolnick, D.; and Tramèr, F. 2024. Stealing Part of a Production Language Model. *ICML*.

Carlson, A. M. 2017. The Need for Transparency in the Age of Predictive Sentencing Algorithms. *Iowa Law Review*, 103.

Carvalho, T.; Moniz, N.; Faria, P.; and Antunes, L. 2023. Towards a Data Privacy-Predictive Performance Trade-off. *Expert Systems with Applications*, 223.

Cattell, S.; Ghosh, A.; and Kaffee, L.-A. 2024. Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities. *AAAI/ACM Conf. on AI, Ethics, and Society (AIES)*.

Cavallar, S.; Dodson, B.; Lenstra, A. K.; Lioen, W.; Montgomery, P. L.; Murphy, B.; te Riele, H.; Aardal, K.; Gilchrist, J.; Guillerm, G.; Leyland, P.; Marchand, J.; Morain, F.; Muffett, A.; Putnam, C.; Putnam, C.; and Zimmermann, P. 2000. Factorization of a 512-Bit RSA Modulus. 1–18. EUROCRYPT.

CERT. 2024. CERT Guide to Coordinated Vulnerability Disclosure. https://certcc.github.io/CERT-Guide-to-CVD/topics/phases/deployment/.

Chakraborty, R. S.; and Bhunia, S. 2009. Security through obscurity: An approach for protecting Register Transfer Level hardware IP. IEEE International Workshop on Hardware-Oriented Security and Trust.

Cooper, A. F.; Moss, E.; Laufer, B.; and Nissenbaum, H. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. *ACM Conf. on Fairness, Accountability, and Transparency*, 864–876.

Cristofaro, E. D. 2024. Synthetic Data: Methods, Use Cases, and Risks. *IEEE Secur. Priv.*, 22(3): 62–67.

Culp, S. 2001. It's Time to End Information Anarchy. Microsoft Technet essay.

Curtmola, R.; Garay, J. A.; Kamara, S.; and Ostrovsky, R. 2006. Searchable symmetric encryption: improved definitions and efficient constructions. 79–88. ACM Conf. on Computer and Communications Security (CCS).

Dev, J.; Akhuseyinoglu, N.; Kayas, G.; Rashidi, B.; and Garg, V. 2024. Building Guardrails in AI Systems with Threat Modeling. *Digital Government: Research and Practice*.

Diffie, W. 2003. Risky Business: Keeping Security a Secret. ZDNet.

Dwork, C. 2006. Differential Privacy. *International Conf. on Automata, Languages, and Programming (ICALP)*, 1–12.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference '06*, 265–284. Theory of Cryptography Conference (TCC).

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9: 211–407.

et al., J. F. 2022. Piloting a survey-based assessment of transparency and trustworthiness with three medical AI tools. *Healthcare (Basel)*.

Fehr, J.; Citro, B.; Malpani, R.; Lippert, C.; and Madai, V. I. 2024. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6.

Felzmann, H.; Fosch-Villaronga, E.; Lutz, C.; and Tamò-Larrieux, A. 2020. Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, 26: 3333–3361.

Ferrara, E. 2023. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6.

Fine, A.; and Marsh, S. 2024. Judicial leadership matters (yet again): the association between judge and public trust

for artificial intelligence in courts. *Discover Artificial Intelligence*, 4.

Foundation, E. F. 2008. MBTA v. Anderson. www.eff.org/cases/mbta-v-anderson.

Foundation, E. F. 2013. U.S. v. Auernheimer. www.eff.org/cases/us-v-auernheimer.

Fukawa, N.; Zhang, Y.; and Erevelles, S. 2021. Dynamic Capability and Open-Source Strategy in the Age of Digital Transformation. *Journal of Open Innovation: Technology, Market, and Complexity*, 7: 175.

Gade, P.; Lermen, S.; Rogers-Smith, C.; and Ladish, J. 2024. BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B. *arXiv.2311.00117*.

Gamero-Garrido, A.; Savage, S.; Levchenko, K.; and Snoeren, A. C. 2017. Quantifying the Pressure of Legal Risks on Third-party Vulnerability Research. *ACM Conf. on Computers and Communications Security*, 1501 – 1513.

Gibney, E. 2024. Not all 'open source' AI models are actually open: here's a ranking.

Gilbertson, A.; and Reisner, A. 2024. Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI. Proof News, Wired.

Girgvliani, S. 2023. Public Procurement Transparency and its Potential to Reduce Corruption in Low-Income Countries. *CSIS - Center for Strategic and International Studies*.

Goodhart, C. 1975. Problems of monetary management: the U.K. experience. *Papers in monetary economics*, 1–20.

Graziani, M.; Dutkiewicz, L.; Calvaresi, D.; Amorim, J. P.; Yordanova, K.; Vered, M.; Nair, R.; Abreu, P. H.; Blanke, T.; Pulignano, V.; Prior, J. O.; Lauwaert, L.; Reijers, W.; Depeursinge, A.; Andrearczyk, V.; and Müller, H. 2022. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, 56: 3473–3504.

Gu, X.; Tianqing, Z.; Li, J.; Zhang, T.; Ren, W.; and Choo, K.-K. R. 2022. Privacy, accuracy, and model fairness tradeoffs in federated learning. *Computers and Security*, 122.

Górriz, J.; Álvarez Illán, I.; Álvarez Marquina, A.; Arco, J.; Atzmueller, M.; Mallarini, F.; Barakova, E.; Bologna, G.; Bonomini, P.; Castellanos-Dominguez, G.; Castillo-Barnes, D.; Cho, S.; Contreras, R.; Cuadra, J.; Domínguez, E.; Domínguez-Mateos, F.; Duro, R.; Elizondo, D.; Fernández-Caballero, A.; Fernandez-Jover, E.; and Ferrández-Vicente, J. 2023. Computational approaches to Explainable Artificial Intelligence: Advances in theory, applications and trends. *Information Fusion*, 100.

Hind, M.; Houde, S.; Martino, J.; Mojislovic, A.; Piorkowski, D.; and Richard, J. e. a. 2020. Experiences with improving the transparency of AI models and services. *Conf. on Human Factors in Computing Systems*, 1–8.

Hosanagar, K. 2024. Regulating AI: Getting the Balance Right. *Knowledge at Wharton*.

in the Office of Technology, S. 2024a. On Open-Weights Foundation Models. U.S. Federal Trade Commission.

in the Office of Technology, S. 2024b. P=NP? Not exactly, but here are some research questions from the Office of Technology. U.S. Federal Trade Commission.

Jain, A.; Maleki, A.; and Saade, N. 2024. Methods for Adapting Large Language Models. Technical report, Meta AI.

Jones, D. W. 2007. Computer Security Versus the Public's Right to Know. *Computers, Freedom and Privacy*.

Jumper, J.; Richard Evans, A. P.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; and et al., M. Z. 2021. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596: 583–589.

Kahn, D. 1996. *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*. Scribner. ISBN 0684831309.

Katz, J.; and Lindell, Y. 2015. *Introduction to Modern Cryptography*. CRC Press, second edition.

Kerckhoffs, A. 1883. La Cryptographie Militaire. In *Journal des sciences militaires*, volume 9, 5–38, 161–191.

Khan, B.; Fatima, H.; Qureshi, A.; Kumar, S.; Hanan, A.; Hussain, J.; and Abdullah, S. 2023. Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. *Biomedical Materials and Devices*.

Liu, Q.; Safavi-Naini, R.; and Sheppard, N. P. 2003. Digital rights management for content distribution. In *ACM ACSW Frontiers '03*, 49–58. ACM ACSW Frontiers.

Longpre, S.; Kapoor, S.; Klyman, K.; Ramaswami, A.; Bommasani, R.; Blili-Hamelin, B.; Huang, Y.; Skowron, A.; Yong, Z.-X.; Kotha, S.; Zeng, Y.; Shi, W.; Yang, X.; Southen, R.; Robey, A.; Chao, P.; Yang, D.; Jia, R.; Kang, D.; Pentland, S.; Narayanan, A.; Liang, P.; and Henderson, P. 2024. A Safe Harbor for AI Evaluation and Red Teaming. *Knight First Amendment Institute*.

Mercuri, R. T.; and Neumann, P. G. 2003. Security by Obscurity. *Communications of the ACM*, 46.

Monteiro, W. R.; and Reynoso-Meza, G. 2023. A multiobjective optimization design to generate surrogate machine learning models in explainable artificial intelligence applications. *EURO Journal on Decision Processes*, 11.

Moura, G. C. M.; and Heidemann, J. 2023. Vulnerability Disclosure Considered Stressful. In *SIGCOMM Computer Communication Review*, volume 53, 2–10. ACM SIGCOMM Computer Communication Review.

Narayanan, A.; and Shmatikov, V. 2008. Robust Deanonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy*. IEEE Symposium on Security and Privacy.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; and Lee, K. 2023. Scalable Extraction of Training Data from (Production) Language Models. *arXiv.2311.17035*.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy*. IEEE Symposium on Security and Privacy.

NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

NIST. 2024. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST.

Novelli, C.; Taddeo, M.; and Floridi, L. 2023. Accountability in artificial intelligence: what it is and how it works. *AI and Society*.

Ohm, P. 2009. Broken Promises of Privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57.

Park, S.; and Albert, K. 2024. A Researcher's Guide to Some Legal Risks of Security Research.

Patel, F.; and Toomey, P. C. 2024. Bringing Transparency to National Security Uses of Artificial Intelligence. *justsecurity.org*.

Polemi, N.; Praça, I.; Kioskli, K.; and Bécue, A. 2024. Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. *Frontiers in Big Data*.

Raimondo, G. M.; and Locascio, L. E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, National Institute of Standards and Technology.

Rajpurkar, P.; Chen, E.; Banerjee, O.; and Topol, E. J. 2022. AI in health and medicine. *Nature Medicine*, 28: 31–38.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-Teaming the Stable Diffusion Safety Filter. *Artificial Intelligence*.

Rauch, B. 2022. Operation Charlie: Hacking the MBTA CharlieCard from 2008 to Present. *Bobbyrsec-Medium*.

Raymond, E. S. 1999. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly and Associates. ISBN 1-565-92724-9.

Rubinstein, I. S.; and Hartzog, W. 2016. Anonymization and Risk. *Washington Law Review*, 91.

Schneier, B. 2003. *Beyond Fear: Thinking Sensibly About Security in an Uncertain World*. Springer.

Schneier, B. 2004. The Non-Security of Secrecy. *Communications of the ACM*, 47.

Schneier, B. 2023. Trustworty AI Means Public AI [Last Word]. *IEEE Security and Privacy*, 21: 95–96.

S.D.N.Y., U. D. C. 2023. The New York Times Company v. Microsoft Corporation et al., No. 1:23-cv-11195 (S.D. N.Y. 2023).

Shipman, A. 2019. Don't Be Afraid to Code in the Open: Here's How to Do It Securely. UK Technology Blog.

Specter, M. A.; Koppel, J.; and Weitzner, D. J. 2020. The Ballot is Busted Before the Blockchain: A Security Analysis of Voatz, the First Internet Voting Application Used in U.S. Federal Elections. In *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, 1535–1553. USENIX Security Symposium.

Stadler, T.; Kulynych, B.; Gastpar, M. C.; Papernot, N.; and Troncoso, C. 2024. The Fundamental Limits of Least-Privilege Learning. *arXiv*.

Suresh, H.; and Guttag, J. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms Mechanisms, and Optimization*. ACM.

Sweeney, L. 2000. Simple Demographics Often Identify People Uniquely. *Carnegie Mellon University, Data Privacy Working Paper*.

Thomas, D. R.; Pastrana, S.; Hutchings, A.; Clayton, R.; and Beresford, A. R. 2017. Ethical issues in research using datasets of illicit origin. In *Internet Measurement Conference*. IMC '17: Proceedings of the 2017 Internet Measurement Conference.

Tiku, N. 2024. Top AI researchers say OpenAI, Meta and more hinder independent evaluations. Washington Post.

Tohe, L. 2022. Code Talkers Were America's Secret Weapon in World War II. Humanities: The Magazine of the NEH.

Union, E. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council. *Official Journal of the European Union*.

Vagle, J. 2020. Cybersecurity and Moral Hazard. *Stanford Technology Law Review*.

Vestager, M. 2024. Speech at the European Commission workshop on "Competition in Virtual Worlds and Generative AI".

Viega, J.; and McGraw, G. 2001. *Building Secure Software: How to Avoid Security Problems the Right Way*. Addison-Wesley Professional. ISBN 9780672334092.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv:2112.04359*.

White, M.; Haddad, I.; Osborne, C.; Yanglet, X.-Y. L.; Abdelmonsef, A.; and Varghese, S. 2024. The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence. arXiv.2403.13784.

Wicker, S. B. 2021. The Ethics of Zero-Day Exploits: The NSA Meets the Trolley Car. *Communications of the ACM*, 64.

Xian, R.; Li, Q.; Kamath, G.; and Zhao, H. 2024. Differentially Private Post-Processing for Fair Regression. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yu, D.; Naik, S.; Backurs, A.; Gopi, S.; Inan, H. A.; Kamath, G.; Kulkarni, J.; Lee, Y. T.; Manoel, A.; Wutschitz, L.; Yekhanin, S.; and Zhang, H. 2024. Differentially Private Fine-tuning of Language Models. *J. Priv. Confidentiality*, 14(2).

Zouridis, S.; and Bovens, M. V. E. M. 2019. *Discretion and the quest for controlled freedom*, chapter Automated Discretion, 313–329. Palgrave Macmillan.

Zuckerberg, M. 2024. Open Source AI Is the Path Forward. Meta Newsroom.