Today, a significant portion of mission-critical work traditionally done by humans (e.g., driving cars, approving loans, medical triaging) is on the verge of being replaced by machine learning (ML). Historically, not considering human interactions with conventional software systems has led to significant harm[1,2,3]. This is even more true for emerging ML systems as there is a lack of principled methods to construct safe and secure human-ML interaction paradigms. To prevent similar harm in ML-based systems, it is paramount that we understand vulnerabilities and apply safeguards now, while they are being designed and deployed.

My work discovers how human interaction impacts ML security in two ways: How human factors can be 1) exploited to reduce security and 2) harnessed to improve security. Since ML-enabled abuse is becoming increasingly common, I investigate how lay users perceive and react to new attacks, e.g., how social media users react to deepfakes [USENIX Sec. 2022, Revision 2024a]. As ML is beginning to be applied in security-critical systems, I evaluate how usable these tools are for technical users, e.g., how easy it is for ML developers to apply security defenses [CHI 2022, USENIX Sec. 2023, IEEE SP 2023a]. Additionally, I leverage my background in system security [ACSAC 2020, ACSAC 2022, IEEE SP 2023b] to create usable solutions for highly technical end users. Lastly, to promote well-grounded results in usable security research, I analyze and improve human subject methodology [WWW 2022, Revision 2024b].

My work uses both social science methods and software evaluation techniques. In human-subject work, I learn user perceptions with observational techniques (e.g., interviews/surveys) and discover how factors influence security with controlled experiments. These vary in scope from 90-minute interviews for critically understanding individual experiences, to thousands of measured observations for discovering generalizable trends. For software, I use benchmarking techniques to rigorously compare systems. Combined, I holistically assess software security via technical and usable metrics.

My research has directly impacted industry and academia. The vulnerabilities I discovered in fitness privacy-obfuscation mechanisms have led to user vulnerability disclosures on the Strava fitness app[4]. My work on deepfakes attracted interest from LinkedIn which subsequently employed new deepfake defenses[5]. My proposed metrics for quantifying audit log security have been recommended by others to formalize comparisons across literature[6]. My expertise led me to consult for "Partnership on AI", an NGO composed of a diverse set of organizations (including Meta, ACLU, The New York Times, ACM) that fosters responsible development of AI systems. My work has been covered by a variety of news outlets (New Scientist, The Transmitter, The 21st Show) and made into educational content (Futurum Careers). I have also been awarded the NSF GRFP for my work on [IEEE SP 2023a].

Looking forward, *I intend to understand how humans interact with emergent ML systems in three areas*: whether provenance-based mitigations to deepfakes are meaningful to users, how large language models are (un)safely used by open-source and abuser communities, and whether new human-AI interaction paradigms for incident response are practical.

Main Topic 1: ML-Enabled Abuse

ML's ability to generate images and text that are indistinguishable from reality allows adversaries to craft "deepfake" personas at scale. These personas represent a pressing societal concern. Deepfakes can be used to attack social media users via spear-phishing and misinformation; in the US, concerns over deepfakes' effect on public opinion and elections[7] resulted in mitigatory executive orders[8]. While companies moderate this content, many deepfake profiles remain undetected and require end-users to filter content themselves. To understand the effects of deepfakes-based abuse on social media, my research investigates how lay users

[1] Leveson, N. et al. "An investigation of the Therac-25 accidents: The operator interface". Computer. 1993.
[2] Salmon, P. et al. "Pilot error versus sociotechnical systems failure: a distributed situation awareness analysis of Air France 447". Theoretical Issues in Ergonomics Science. 2016.
[3] Meshkati, N. "Human factors in large-scale technological systems' accidents: Three Mile Island, Bhopal, Chernobyl". Industrial Crisis Quarterly. 1991.
[4] Meg. "Edit Map Visibility, 2023." Strava Support. 2023.
[5] Rodriguez, O. "New LinkedIn profile features help verify identity, detect and remove fake accounts, boost authenticity". LinkedIn Official Blog. 2022.
[6] Zipperle, M. et al. "Provenance-based intrusion detection systems: A survey". ACM Computing Surveys. 2022.
[7] O' Brien, M. "Meta and X questioned by lawmakers over lack of rules against AI-generated political deepfakes". AP News. 2023.
[8] "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence". The White House. 2023.

perceive and react to deepfake-generated profiles [USENIX Sec. 2022] and whether the moderation process of deepfakes results in disproportionate harm to real users based on their identity [Revision 2024a].

How Do Lay Users Respond to Deepfake Personas? [USENIX Sec. 2022] My work is the first to perform a mixed-methods survey to assess whether users accept connection requests from social media profiles containing deepfake text and images. My evaluation of deepfake-enabled social engineering differs from prior work which has only investigated individual media (e.g., text or photos exclusively) and provided explicit deepfake searching tasks to participants. I found several surprising results. Even in the most favorable conditions with the most noticeable deepfake artifacts (e.g., malformed faces), many participants still connected to deepfake profiles. This tells us that deepfake quality is not directly proportional to attacker success. *By the time profiles get to users, it's generally too late to prevent harm.*

To investigate whether deepfake education mitigates harm, I warned a subset of users what artifacts typically look like. Surprisingly, this was not only ineffective but actively harmful. I found that deepfake warnings led to antisocial behaviors against other real profiles. In searching for deepfakes, users began misattributing real content as artifact-containing. Worse, I found that the perceptions of artifacts are sometimes informed by racial and gender stereotypes. Thus, given the low efficacy and potential discriminatory harms, I found a unique case in which user warnings broadly harm users. As a result, I recommend that platforms leave deepfake moderation to automatic systems and professional human teams. However, these results led us to question whether human moderation also suffers from identity-based biases; this directly informed my next study.

Does Deepfake Moderation Lead to Biased User Harm? [Revision 2024a] This mixed-method user study is the first to directly investigate whether certain gender or racial identities are disproportionately categorized as artificial during moderation. To do this, users are shown a set of profiles and asked to select the deepfakes; however, all shown profiles are real and across various identities. I found statistical evidence of biases during the moderation of artificial content. Specifically, profiles of Black women and Black men are found to have significantly less perceived artificiality compared to profiles of white women and white men. I also found that users who share the same identity as the assessed profile are significantly less likely to misclassify them. Surprisingly, these biases are not just due to stereotypes, but also from upstream perceptions of algorithmic biases: since ML models poorly represent Black women and men, users believe that these identities are less likely to be deepfaked. I suggest several mitigations to prevent bias during content moderation such as having a diverse set of moderators and UI designs that deemphasize identity-laden profile fields, like images and names.

Main Topic 2: ML and Security

ML has been implemented in a variety of security-sensitive applications; however, little consideration has been given to its security implications. To address this gap, my work investigates whether security analysts believe that ML-security tools are beneficial or detrimental to accomplishing their tasks compared to traditional tools [IEEE SP 2023a], and what sociotechnical barriers prevent ML developers from implementing adversarial ML defenses [USENIX Sec. 2023].

Do Security Analysts Find Benefits in ML Tools? [IEEE SP 2023a] ML is becoming increasingly integrated in security analysts' tools for detecting and mitigating security events; however, no work has asked security analysts whether ML tools are even beneficial. I answer this by interviewing security practitioners on the benefits and pain points ML-based tools provide. I also ask whether techniques explaining the decisions of ML models (ML explanations) alleviate any concerns.

I found that despite the security community's focus on ML explanations, practitioners report that the (in)ability of ML tools to correctly classify events is still the most influential factor for practitioners' tool choice. Generally, practitioners felt that ML's ability to detect novel attacks is overshadowed by its disruptive false alerts. Thus, while extra features are nice, ML use is still limited due to its low efficacy. Practitioners also noted that explanations would be helpful, but in ways not typically considered by the security community; for instance, as teaching devices for understanding security events and training new

employees. Practitioners proposed several improvements for explanations for such purposes: these included providing direct remedial advice and additional contextualization. Thus, this work helps redefine the research agenda for ML-based security tools by refocusing efforts on efficacy (e.g., correct classification) and providing concrete improvements for ML explanations.

Why Are Adversarial ML Defenses Not Applied in Industry? [USENIX Sec. 2023] Attackers have begun performing adversarial ML (AML) attacks to steal tens of millions of dollars and poison malware classification systems[9,10]; however, companies do not implement mitigations[11]. While this phenomenon has been previously attributed to a lack of awareness, its cause is not understood. By interviewing machine learning practitioners, my work takes a broader perspective and investigates root sociotechnical causes for the disconnect between the need and implementation of AML defenses in industry.

I found several phenomena unique to ML that heighten the difficulty of defense implementations. Culturally, security was perceived as detached from development: even when aware of model-related vulnerabilities, several ML practitioners felt that security concerns should have no bearing on model development. Organizationally, company goals may conflict and prevent defenses: some practitioners noted that accuracy-decreasing defenses would not be used in models that were financially important. My insight is that these issues are extremely analogous to those faced and mitigated in software security. As such, I recommend the use of proven solutions to accommodate these sociotechnical constraints: e.g., cultural changes via the promotion of security champions and a greater focus on cost-aware security mechanisms. Thus, this work establishes, that beyond the technical issues the field of ML security often focuses on, the sociotechnical issues faced by ML practitioners require consideration.

Other Topics and Contributions

Alongside my primary work in human factors in ML, my work in system security and HCI methodology better equips me to develop usable security solutions for technical- and lay-end users. In the field of audit log security, my work building novel log reduction algorithms [ACSAC 2020], producing the first open-source log reduction framework [ACSAC 2022], and systemizing current knowledge [IEEE SP 2023b] allows me to formally assess software, and build solutions for technical-end users. My work improving the methodology of online recruitment of participants [WWW 2022] and the analysis/interpretation of sociodemographics in the context of security behaviors [Revision 2024b] helps researchers, including myself, ensure that study results are well-grounded.

Future Work

Looking forward, I intend to understand how humans interact with emergent ML systems in three areas of immediate impact: provenance-based mitigations to deepfakes, safety practices of large language models (LLMs) in open-source communities, and misuse of LLMs by bad actors. Additionally, I will pursue a longer-term research agenda that investigates a new human-AI interaction paradigm for incident response.

Short Term: Can Provenance Mitigate Deepfakes and Misinformation? A promising direction to reduce online deception is media provenance[12] which *proves the origin of media*; however, there is little research on whether users *understand* provenance, or whether it would cause *additional harm*. First, people hold biases when viewing politically opposing/supporting media; I will investigate whether cryptographically verified provenance diminishes this bias. Second, given the importance of peer-to-peer news in online social networks, a decentralized trust model (e.g., a web of trust) may be a fitting paradigm for provenance; I will investigate how understandable and useful decentralized trust models are to lay users, and whether they may result in harmful side-effects, such as echo chambers. Third, there exists a fundamental tension between provenance and privacy; I will investigate if the use of provenance causes

[9]McAfee Advanced Threat Research. "VirusTotal Poisoning". MITRE ATLAS. 2020.
[10] Borak, M. "Chinese government-run facial recognition system hacked by tax fraudsters: report". South China Morning Post. 2021.
[11] Kumar, R. et al. "Adversarial machine learning-industry perspectives". In Proc. of IEEE SP Workshops, 2020.
[12] "C2PA Specifications". The Coalition for Content Provenance and Authenticity. 2023.

unintentional harm to marginalized groups (e.g., whistle-blowers, and political dissidents). I will understand what concerns these users may have, what changes may alleviate them, and whether the choice to not use provenance will cause viewers to disbelieve content from these groups.

Short Term: Does The Open-Source Community Consider LLM Security, Privacy, or Safety? The open-source community has found value in LLMs, powering applications used by tens of thousands of people for financial reporting[13] and code completion[14]. Unfortunately, LLMs are also susceptible to adversarial exploits (e.g., jailbreaking, prompt injection) and unintentional faults (e.g., hallucinations). While mitigations exist, *it is unknown to what extent the open-source community understands these risks or mitigates them.* I intend to discover this via a combination of community measurements and interviews. First, to discover the risk surface, I will crawl for open-source projects in LLM-developer communities (e.g., dev.to, llmops.space). For each project, I will then analyze whether a security/privacy-sensitive use case exists, and if so, whether any warnings/guardrails are implemented. Second, to understand community perception of LLM risks, I will see if communities recommend/enforce safety guidelines and how often discussions of risk emerge among members. Third, via a set of interviews, I will learn how LLM developers and community admins perceive these threats and the barriers toward mitigation. This will inform an understanding of the existing gaps towards LLM safety, and how improvements in model distribution, interfaces, and community administration can better achieve responsible LLM development.

Short Term: How Do Abusers Misuse LLMs? The open-sourcing and jail-breaking vulnerabilities of LLMs have sparked concerns over malicious use; however, *there is little understanding of whether or how these LLMs are misused in practice*. Using underground forums focused on selling elicit security/abuse tools, I will evaluate the prevalence, use cases, and technical set-ups for LLM-based abuse. I also will reach out for anonymous interviews with LLM-abuse developers to better understand the motivations, technical background, and development efforts required to produce these tools. This will inform how abuse-oriented LLMs are developed and distributed in order to prevent such actions.

Long Term: Is a Human-Supported AI Defense Practical? Current incident response workflows support a human-driven paradigm: from alert collection to remediation, humans are the decision-makers, and AI tools are invoked to inform humans or apply operations. However, as security experts are in limited supply, security operation centers are quickly overwhelmed. My question is *should humans drive the workflow and invoke AI*, or *should AI agents drive the workflow and invoke humans*? I propose investigating a human-supported AI defense paradigm in which AI agents autonomously manage the overarching security response, only prompting human analysts when required (e.g., for creative or risk-heavy tasks). Unlike Git Security Co-pilot which acts as an analyst tool, this paradigm would shift the bulk of the work off of humans, relaxing a major resource constraint of enterprises.

Investigating this paradigm requires a series of questions to be answered: *First*, what portions of the incident response workflow are most appropriate to remain human-led and which can be safely handed to AI agents? *Second*, given a segmentation of responsibility, how should AI agents interact with analysts to invoke their help; what context and information is needed for security analysts to perform their tasks? *Third*, what safety guardrails or insights into AI agents do humans need to feel assured that security-critical operations are being performed? *Fourth*, can an AI agent be developed that meets the technical and usable requirements needed for this paradigm shift? Beyond alleviating one of the most constrained resources in security (i.e., human labor), this research agenda will investigate a new human-AI paradigm in security, only made possible due to recent advances in AI agents. Given my background in human factors in ML security, security analysts' perception of ML-enabled tools, and my design of system security algorithms and frameworks, I am extremely well-suited to investigate this line of work.

[13] Gowda, V. "How I built the Streamlit LLM Hackathon winning app — FinSight using LlamaIndex". LlamaIndex Blog. 2023.
[14] FauxPilot. "FauxPilot". Github Repository. 2023.

References

[USENIX Sec. 2023] **Jaron Mink**, Harjot Kaur, Juliane Schmüser, Sascha Fahl, Yasemin Acar. ""Security is not my field, I'm a stats guy": A Qualitative Root Cause Analysis of Barriers to Adversarial Machine Learning Defenses in Industry". In *Proc. of USENIX Security*, 2023.

[IEEE SP 2023a] **Jaron Mink**, Hadjer Benkraouda, Limin Yang, Arridhana Ciptadi, Ali Ahmadzadeh, Daniel Votipka, Gang Wang. "Everybody's Got ML, Tell Me What Else You Have: Practitioners' Perception of ML-Based Security Tools and Explanations". In *Proc. of IEEE S&P*, 2023.

[IEEE SP 2023b] Muhammad Adil Inam, Yinfang Chen, Noor Michael, Jason Liu, **Jaron Mink**, Sneha Gaur, Adam Bates and Wajih Ul Hassan. "SoK: History is a Vast Early Warning System: Auditing the Provenance of System Intrusions". In *Proc. of IEEE S&P*, 2023.

[USENIX Sec. 2022] **Jaron Mink**, Licheng Luo, Natã M. Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. "DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks". In *Proc. of USENIX Security*, 2022.

[CHI 2022] **Jaron Mink**, Amanda Rose Yuile, Uma Pal, Adam J Aviv, and Adam Bates. "Users Can Deduce Sensitive Locations Protected by Privacy Zones on Fitness Tracking Apps". In *Proc. of CHI*, 2022.

[WWW 2022] Ziyi Zhang, Shuofei Zhu, **Jaron Mink**, Aiping Xiong, Linhai Song and Gang Wang. "Beyond Bot Detection: Combating Fraudulent Online Survey Takers". In *Proc. of The Web Conference*, 2022.

[ACSAC 2022] Muhammad Adil Inam, Akul Goyal, Jason Liu, **Jaron Mink**, Noor Michael, Sneha Gaur, Adam Bates, Wajih Ul Hassan. "FAuST: Striking a Bargain between Forensic Auditing's Security and Throughput". In *Proc. of ACSAC*, 2022.

[ACSAC 2020] Noor Michael, **Jaron Mink**, Jason Liu, Sneha Gaur, Wajih Ul Hassan, and Adam Bates. "On the Forensic Validity of Approximated Audit Logs". In *Proc. of ACSAC*, 2020.

[Revision 2024a] **Jaron Mink**, Miranda Wei, Collins W. Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M. Redmiles, Gang Wang. "It's Trying Too Hard To Look Real: Deepfakes Moderation Mistakes and Identity-Based Bias". *Under Revision at CHI*, 2024.

[Revision 2024b] Miranda Wei, **Jaron Mink**, Tadayoshi Kohno, Elissa M. Redmiles, Franziska Roesner. "SoK or So(L)K? On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors". *Under Revision at USENIX Security*, 2024.