Keeping Humans in Practice: Skill Atrophy as a Design Constraint for Human-AI Teaming in Cybersecurity

Tarini Saka
Max Planck Institute for Security and Privacy

Abstract

Human–AI collaboration is becoming central to cybersecurity operations. These systems promise speed and scalability, but also create an oversight paradox: humans are expected to supervise AI-assisted workflows, while the same workflows may reduce their opportunities to practice the skills needed for meaningful intervention. This creates a risk of *skill atrophy*, whose consequences are especially acute in a high-stakes domain such as cybersecurity. Analysts need to retain the skills to intervene when automation fails; otherwise, errors can escalate into serious security incidents. This position paper advocates for an atrophy-aware lens for human–AI security teaming, centered on competency risk: the risk that AI assistance reduces practice of skills humans must retain for effective oversight. The paper proposes preliminary design mechanisms including competency-risk task classification, recurring AI-off and AI-degraded drills, and reflection tools that make AI reliance visible. Treating skill retention as an operational requirement can help ensure that AI improves cybersecurity work without eroding the human expertise.

1 Introduction: The Oversight Paradox

Human–AI collaboration is becoming central to cybersecurity work. AI systems can summarize alerts, correlate telemetry, explain commands, draft incident reports, and recommend next steps [15, 25]. Agentic systems may soon be able to query infrastructure, modify configurations, open tickets, run playbooks, and coordinate response actions [9, 27]. This automation is often presented as a way to reduce heavy workloads and help deal with a shortage of cybersecurity experts.

Workshop on Designing Human–AI Teaming for Security (HATS), co-located with the Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23, 2026, Hannover, Germany.

Yet this framing hides a tension. The more AI systems perform routine and cognitively demanding parts of security work, the fewer opportunities humans may have to practice the skills needed to supervise those systems. This is especially problematic because human oversight is still expected precisely when AI fails: erroneous outcomes, during novel attacks, ambiguous evidence, or adversarial manipulation [6, 21, 23]. A "human-in-the-loop" may not be able to effectively oversee security workflows if they lose the ability to reason through tasks without external assistance. Another key concern is *automation surprise*, when an operator's understanding of an automated system diverges from its actual state or actions, leaving the operator unsure how the system reached its current mode or how to intervene [24]. Skill atrophy may compound this problem: analysts may lack both the practiced ability to take over manually and the situational awareness needed to reconstruct what the AI has done. This critical dimension of human–AI collaboration remains understudied and should receive greater attention in cybersecurity research and discourse.

This position paper examines the risk of skill atrophy arising from such sustained reliance on AI systems [20]. *Skill atrophy* refers to the gradual weakening of a person's ability to perform a task because they practice it less often. This has become a growing concern across domains as AI systems are increasingly used to support or replace cognitive work [7, 8, 22]. In AI-assisted cybersecurity workflows, skill atrophy may occur when analysts repeatedly delegate parts of security reasoning, such as interpreting alerts, generating hypotheses, tracing evidence, or planning responses, to AI systems. Over time, humans may remain responsible for oversight while receiving fewer opportunities to practice the required skills. This paper argues that skill atrophy should therefore be treated as a key constraint for sustainable human–AI teaming in cybersecurity. The central position is simple: *critical skill retention is a requirement for meaningful human oversight*. I propose a *atrophy-aware* lens for designing security workflows that preserve human competence while still taking advantage of AI-enabled automation.

2 Skill Atrophy as a Security Risk

The risk that automation can weaken human performance is not new. Classic work shows that automation can leave humans responsible for failures while reducing their routine involvement and situation awareness [3, 6]. People may also over-rely on automation, avoid it, or be placed in poorly designed roles [17]; these patterns can lead to complacency and automation bias, where users monitor less carefully or accept recommendations too readily [16].

Recent work suggests that these concerns extend to AI-assisted work. Rinta-Kahila et al. show how cognitive automation can cause skill erosion through reliance, complacency, and reduced mindful performance [19]. Macnamara et al. argue that AI assistance can accelerate skill decay among experts and hinder skill acquisition among learners, while also making these effects difficult for users to notice [12]. Generative AI intensifies this concern because it does not merely retrieve information; it produces plausible analyses, explanations, code, and recommendations [22]. Lee et al. found that GenAI use can reduce perceived critical-thinking effort, especially when users are confident in the system [11], while early work on AI-assisted coding and writing suggests a tension between productivity and skill formation [8]. This risk may become more acute as cybersecurity systems become increasingly agentic [9, 27]. As AI systems take on more planning, prioritization, and response tasks, humans may have fewer opportunities to practice the reasoning skills needed to validate, override, or recover from AI-driven actions.

This motivates an important stance: *in security-critical domains, human–AI workflows should be designed not only for efficiency, but also for skill preservation.* The field of cybersecurity is adversarial, uncertain, and constantly evolving. Analysts must often reason from incomplete evidence, generate threat hypotheses, identify deception, trace evidence across systems, and make decisions under time pressure. The skills that are likely to atrophy under heavy AI assistance are also the skills most needed when AI recommendations are wrong, incomplete, or manipulated. If AI systems reduce opportunities to practice these skills, then skill atrophy becomes more than an individual productivity concern; it becomes an operational security risk, making human oversight present in the workflow but weak in practice.

3 Risk Preservation in Human-AI Teaming

Human–AI interaction research has long studied how to make AI systems useful, understandable, controllable, and appropriately integrated into human work. For example, Amershi et al. provide guidelines for supporting users across different stages of interaction, including setting expectations before use, supporting effective interaction during use, enabling correction when systems fail, and adapting appropriately over time [1]. In cybersecurity, related work has begun to classify both the work humans do and the ways AI may support it. The NICE Framework decomposes cybersecurity work into roles, tasks, knowledge, skills, and competency areas [18], while recent SOC-focused frameworks classify human–AI collaboration by autonomy level, trust thresholds, human-in-the-loop roles, and teaming paradigms [13, 15]. Other work maps where AI and LLMs can support security operations, such as alert triage, threat intelligence, vulnerability analysis, anomaly detection, incident response, and analyst communication [4, 25, 26]. Broader human-in-the-loop taxonomies also help describe where and how humans participate in AI systems [10]. Together, these frameworks help answer important questions: what cybersecurity tasks exist, where AI can be applied, how much autonomy AI should have, and how humans should remain involved. These perspectives are useful but incomplete for the problem of skill atrophy. However, they do not ask what human capabilities may weaken when AI repeatedly performs parts of the work. This section therefore introduces *competency risk* as a missing dimension for classifying AI-assisted cybersecurity workflows.

Rather than proposing a new taxonomy, this paper proposes a novel *competency-risk lens for evaluating AI-assisted workflows.* A task should be assessed not only by whether AI can perform it accurately, cheaply, or quickly, but also by whether repeated delegation reduces practice of a critical skill humans must retain for effective oversight, recovery, or threat response. Competency risk therefore depends on how important the human skill is, how often AI performs it, and whether the workflow still gives humans chances to practice it.

4 Designing Atrophy-Aware Teaming

The goal of this paper is not to argue that competency-critical tasks should never be automated because AI assistance can be useful or even necessary to manage scale, speed, and complexity. Rather, the claim is that when such tasks are automated or heavily AI-assisted, workflows should include compensating mechanisms that preserve the underlying human skill. Below are some possible mechanisms that can preserve skill and competency in Human–AI collaborative security workflows.

Classify tasks by competency risk. Security teams can begin by breaking existing workflows into smaller tasks and evaluating each task not only by automation feasibility, but also by competency risk. Table 1 provides a preliminary schema for this evaluation. The goal is not to assign fixed labels to specific cybersecurity tasks, but to help teams ask where each task falls based on two dimensions: how important the underlying human skill is, and how much that skill is delegated to AI. Tasks with low competency criticality and low delegation may require little intervention, while tasks with high competency criticality and high delegation may require stronger human involvement, periodic unaided practice, or other skill-

	Low AI delegation	High AI delegation
Low competency criticality	*Low atrophy concern*: AI use is unlikely to weaken skills needed for oversight or recovery.	*Efficiency-oriented automation*: monitor for errors, but skill-retention risk is limited.
High competency criticality	*Skill-preserving assistance*: AI can support the task while humans still practice core reasoning.	H*igh atrophy concern*: workflows should include compensating mechanisms to preserve human skill.

Table 1: A conceptual schema for assessing competency risk in AI-assisted cybersecurity workflows. Competency risk increases when a task depends on skills humans must retain, and those skills are frequently or fully delegated to AI.

preserving mechanisms. This classification can be layered on top of existing frameworks that already describe cybersecurity work roles, SOC workflows, or AI autonomy levels.

Incorporate recurring AI-off and AI-degraded practice. If organizations expect analysts to intervene when AI systems fail, they must also give analysts recurring opportunities to practice intervening without those systems. Many tasks become less frequently practiced as AI systems take over routine sensemaking, summarization, and recommendation. Without deliberate practice, analysts may remain formally responsible for oversight while losing the ability to provide it.

Security teams already use various exercises, simulations, incident-response drills, and CTF-style challenges. These practices can be adapted into AI-off or AI-degraded exercises. In an *AI-off drill*, analysts complete selected tasks without AI assistance. In an *AI-degraded drill*, analysts may use an AI system that is intentionally incomplete, uncertain, or occasionally wrong, forcing them to detect errors, reconstruct evidence, and decide when to override the system.

These drills should be framed as resilience training with the goal to preserve fallback capacity. Automation research shows that operators may be most needed when automation fails, even though automation can reduce their routine practice [3,6]. AI-off and AI-degraded practice directly address this problem by keeping critical skills exercised, measurable, and available during failure conditions.

Support self-reflection and usage awareness. Skill atrophy may be difficult for analysts to notice in day-to-day work because AI assistance often improves short-term performance [12]. A completed ticket, polished incident summary, or plausible explanation can hide the fact that the analyst practiced less of the underlying reasoning process. Atrophy-aware systems should therefore include lightweight mechanisms for self-reflection and usage awareness. This draws on two ideas from organizational psychology. First, intrinsic motivation matters: practitioners may value technical competence, but workplace pressures around speed and ticket closure can push skill maintenance aside. Prior work shows that external rewards and performance pressures can undermine intrinsic motivation under some conditions [5]. Second, reflection after feedback can improve task performance [2]. In cybersecurity, feedback about AI reliance should therefore prompt structured reflection: Which tasks am I delegating? Which skills do I want to preserve? Where did I accept AI output without reconstructing the evidence?

A concrete implementation could be a periodic "AI reliance reflection" session. Analysts could receive private summaries of how often they used AI assistance, which task categories they used it for, how much time they spent in AI-assisted versus unassisted work, and how often they accepted, revised, or rejected AI suggestions. They could also complete similar tasks with and without AI support to compare speed, accuracy, confidence, and error detection. The goal would not be to shame AI use or discourage appropriate assistance, but to make invisible dependency visible and help analysts choose where they need deliberate practice.

Transparency to support active oversight. Transparency and explainability can help build appropriate trust and confidence in an AI system. They can help analysts understand which evidence informed an AI recommendation and potentially support situational awareness and continued engagement with the task [14]. However, explanations alone should not be assumed to preserve skill or ensure effective oversight [21]. Atrophy-aware systems should therefore pair explanations with active-engagement mechanisms, such as requiring analysts to reconstruct the evidence, form an independent judgment, or identify plausible ways in which the AI's recommendation could be wrong.

5 Conclusion and Future Directions

Human–AI teaming in cybersecurity may improve operations, but efficiency is an incomplete design goal. If AI systems reduce opportunities for humans to practice critical security reasoning, then human oversight may become symbolic: present in the workflow, but weak in practice. This concern is not unique to cybersecurity, but cybersecurity makes it especially consequential because humans are needed most when automation is least reliable: during novel attacks, ambiguous evidence, adversarial manipulation, and high-impact response decisions. Such teaming should therefore be designed not only around trust calibration, appropriate automation, and shared agency, but also around long-term skill retention. Treating skill atrophy as a critical risk can help ensure that humans re-

main capable partners in the loop when their judgment matters most. Several directions can support this goal.

Classify tasks by competency risk: Audit AI-assisted workflows by mapping each task against human-skill criticality and degree of AI delegation. A structured SOC review can flag tasks needing skill-preserving safeguards.

Use AI-off and AI-degraded drills: Run regular drills where analysts perform critical tasks without AI or with flawed AI support. These exercises test error detection, evidence reconstruction, and recovery from automation failures.

Support self-reflection: Provide analysts with private summaries of AI reliance, including task types, usage frequency, and assisted versus unaided performance, to highlight dependencies and guide deliberate practice.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2019.

[2] Frederik Anseel, Filip Lievens, and Eveline Schollaert. Reflection as a strategy to enhance task performance after feedback. *Organizational Behavior and Human Decision Processes*, 110(1):23–35, 2009.

[3] Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983.

[4] Mohan Baruwal Chhetri, Shahroz Tariq, Ronal Singh, Fatemeh Jalalvand, Cécile Paris, and Surya Nepal. Towards human-ai teaming to mitigate alert fatigue in security operations centres. *ACM Transactions on Internet Technology*, 24(3):1–22, 2024.

[5] Edward L. Deci, Richard Koestner, and Richard M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6):627–668, 1999.

[6] Mica R. Endsley and Esin O. Kiris. The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2):381–394, 1995.

[7] Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 2025.

[8] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task, 2025.

[9] Nir Kshetri. Transforming cybersecurity with agentic ai to combat emerging cyber threats. *Telecommunications Policy*, 49(6):102976, 2025.

[10] Konstantinos Lazaros, Aristidis G. Vrahatis, and Sotiris Kotsiantis. Human-in-the-loop artificial intelligence: A systematic review of concepts, methods, and applications. *Entropy*, 28(4), 2026.

[11] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–22. ACM, 2025.

[12] Brooke N. Macnamara, Ibrahim Berber, M. Cenk Cavusoglu, Elizabeth Krupinski, Nikhil Nallapareddy, Nicole E. Nelson, Paul J. Smith, Amy L. Wilson-Delfosse, and Subha Ray. Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications*, 9(46), 2024.

[13] Masike Malatji. Augmented intelligence framework for human–artificial intelligence teaming in cybersecurity. *Human-Centric Intelligent Systems*, 5:151–180, 2025.

[14] Vincent Zibi Mohale and Ibidun Christiana Obagbuwa. A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, 8:1526221, January 2025.

[15] Adnan Mohsin, Helge Janicke, Abdelwahed Ibrahim, Iqbal H. Sarker, and Seyit Camtepe. A unified framework for human-AI collaboration in security operations centers with trusted autonomy, 2025.

[16] Raja Parasuraman and Dietrich H. Manzey. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 2010.

[17] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

[18] Rodney Petersen, Danielle Santos, Karen Wetzel, Matthew Smith, and Gregory Witte. Workforce framework for cybersecurity (NICE framework). NIST Special Publication 800-181 Revision 1, National Institute of Standards and Technology, November 2020.

[19] Tapani Rinta-Kahila, Esko Penttinen, Antti Salovaara, Wael Soliman, and Joona Ruissalo. The vicious circles of skill erosion: A case study of cognitive automation. *Journal of the Association for Information Systems*, 24(5):1378–1412, 2023.

[20] Evan F. Risko and Sam J. Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016.

[21] Neele Roch, Hannah Sievers, Noé Zufferey, and Verena Zimmermann. You know why, but still rely: The impact of explainable AI on trust, task load, and performance in cybersecurity decision-making. In *35th USENIX Security Symposium (USENIX Security 26)*. USENIX Association, August 2026. Accepted paper; prepublication version.

[22] Ioan Roxin. Generative ai: The risk of cognitive atrophy. Polytechnique Insights, July 2025. Interview by Anne Orliac. Accessed 31 May 2026.

[23] Tarini Saka, Kalliopi Vakali, Adam D G Jenkins, Nadin Kokciyan, and Kami Vaniea. Judging phishing under uncertainty: How do users handle inaccurate automated advice? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[24] Nadine B. Sarter and David D. Woods. How in the world did we ever get into that mode? mode error and awareness in supervisory control. *Human Factors*, 37(1):5–19, March 1995.

[25] Ronal Singh, Shahroz Tariq, Fatemeh Jalalvand, Mohan Baruwal Chhetri, Surya Nepal, Cecile Paris, and Martin Lochner. Llms in the soc: An empirical study of human-ai collaboration in security operations centres, 2025.

[26] Siddhant Srinivas, Brandon Kirk, Julissa Zendejas, Michael Espino, Matthew Boskovich, Abdul Bari, Khalil Dajani, and Nabeel Alzahrani. Ai-augmented soc: A survey of llms and agents for security automation. *Journal of Cybersecurity and Privacy*, 5(4), 2025.

[27] Vaishali Vinay. The evolution of agentic ai in cybersecurity: From single llm reasoners to multi-agent systems and autonomous pipelines. In *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC)*, pages 1–8. IEEE, 2026.