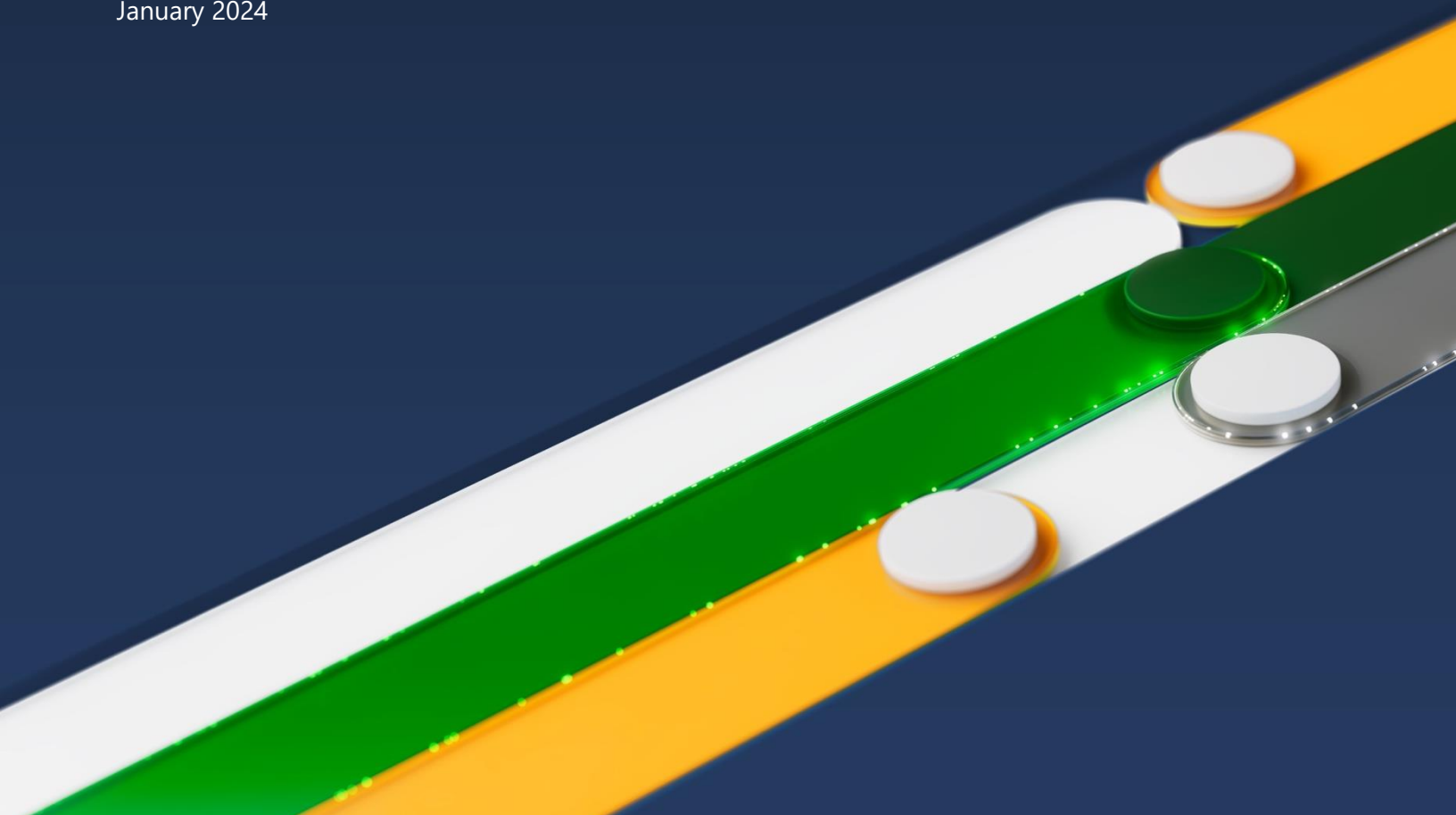
Microsoft Security

Randomized controlled trial for Microsoft Security Copilot

Whitepaper

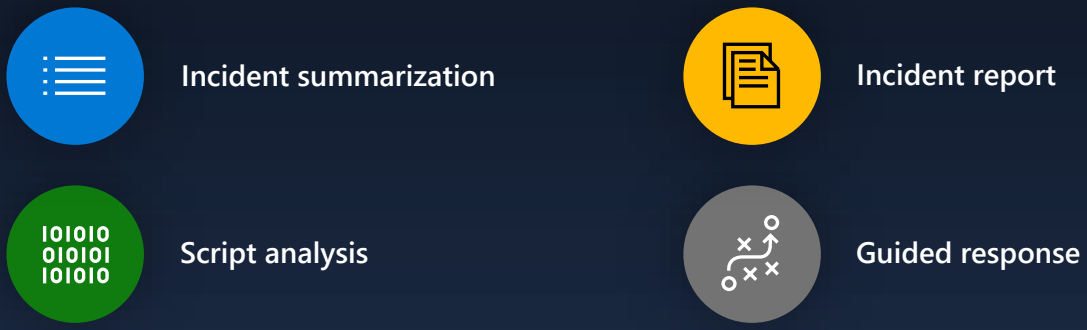
Ben Edelman, James Bono, Sida Peng,
Roberto Rodriguez, Sandra Ho

January 2024



Purpose and methodology

We conducted randomized controlled trials (RCTs) to measure the efficiency gains from using Security Copilot including speed and quality improvements. External experimental subjects logged into a Microsoft Defender XDR (Defender XDR) environment created for this experiment and performed four tasks: incident summarization, script analysis, incident report, and guided response.



We granted half of the subjects ("treatment subjects") access to a standard Defender XDR environment with Microsoft Security Copilot ("Copilot") embedded capabilities. The other half ("control subjects") had the same standard Defender XDR environment without Copilot capabilities. We assigned subjects randomly to groups. Thus, the difference between treatment and control outcomes yields a measurement of the causal impact of Copilot – how we expect outcomes to change if an average control subject uses Copilot.

Our test environment included two sample scenarios. The first was a multistage, hands-on keyboard ransomware attack involving lateral movement, PowerShell script execution, and the use of Microsoft OneNote and Group Policy Objects to distribute payloads. The second was a business email compromise (BEC) financial fraud attack involving a compromised inbox used for lateral movement as well as inbox rule creation, sending suspicious BEC emails, and deleting sent emails.

We provided subjects with an introduction to Defender XDR, then gave them a series of tasks including multiple-choice questions and an incident summary essay. We timed their work in all tasks.

Security professionals

Security novices

We completed two iterations of the test on different subjects: security novices and experienced security professionals. First, in October 2023, we tested security novices. We did not require any special security expertise, and we paid rates commensurate with basic IT skills but not security expertise. This portion of the study thus measures the effect of Copilot on security novices, such as interns and new hires. Second, in December 2023 to January 2024, we tested security professionals who provide security services to large companies (Microsoft and others) through a staffing agency. This second iteration of the study thus measures the effect of Copilot on the seasoned analysts who typically work on security operations for enterprise customers.

Findings - professionals

The findings below are based on our study of 147 security professionals.

Years experience	Analyst level	Subject count
≤2	1	9
3-4	2	28
5-8	3	58
>8	4	52

Accuracy and quality

	Overall	Copilot	Control	P-value
Overall	9.40	9.74	9.07	0.04
Script analysis	3.76	3.97	3.54	0.01
Incident report	3.14	3.13	3.15	0.88
Response	2.96	2.97	2.95	0.17

Findings

Copilot users were **7% more accurate** at the overall task. The difference is statistically significant.

Copilot users were **12% more accurate** at the script analysis task. The difference is statistically significant.

Control users were 1% more accurate at the incident report task.

Copilot users were 1% more accurate at the response task.

Security professionals with Copilot were

7% more accurate

on the three tasks that involved multiple choice questions**[1] [2]

Overall accuracy is on a 15-point score, comprised of three 5-point sections as indicated.

For the incident summarization task, we asked subjects to write incident summaries based on the findings in Defender XDR. Copilot subjects were free to copy-and-paste the incident summary from the Copilot incident summary skill or rewrite as they saw fit. To grade these summaries, we first asked our security experts to identify 15 key facts that should be in an essay about the incident at hand. We then ran an LLM grader to determine which of those key facts were included in each incident summary. Importantly, the LLM grader was not prompted to determine whether the summaries contained ungrounded claims.

The LLM grader found that essays from Copilot users had an average of 6.95 of these facts, versus 4.67 in essays from control users. So, security professionals using Copilot got a 49%** higher content score. Users with Copilot also produced higher quality writing, earning a 10%** higher score in writing quality.

	Overall	Copilot	Control	P-value
Content	5.82	6.95	4.67	0.00
Quality	2.83	2.96	2.70	0.04

Findings

Copilot users got a **49% higher content score** on their essay.
The difference is statistically significant.

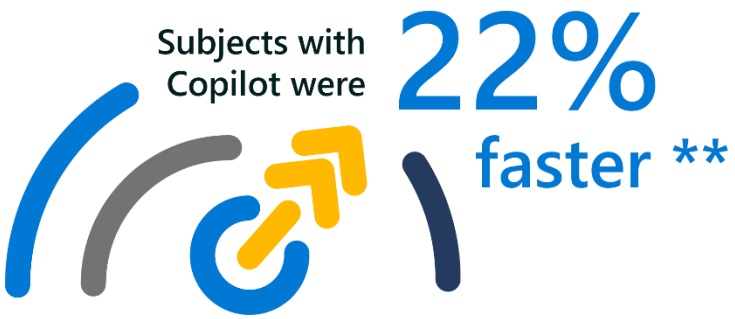
Copilot users got a **10% higher quality score** on their essay.
The difference is statistically significant.

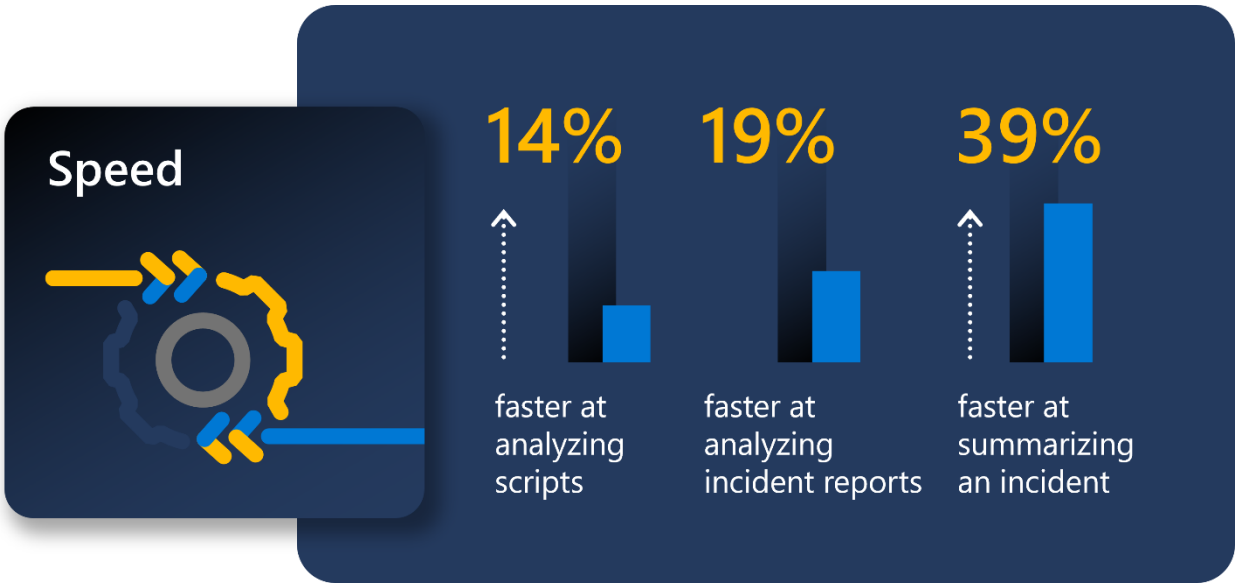
[1] Throughout this section and its counterpart for security novices, we limit our analysis to users who made a good-faith effort. We exclude subjects whose scores are lower than guessing randomly among multiple choice questions.

[2] Throughout, * denotes statistical significance at P<0.10, and ** denotes significance at P<0.05.

Speed

Overall, security professionals with Copilot finished the overall tasks 22%** faster. This finding resulted in part from them finishing incident summarization 39%** faster, analyzing scripts 14%** faster, and analyzing incident reports 19% faster**.





We reach this finding via a linear regression framework that proceeds in three steps. First, we estimate the task duration as a function of accuracy for the control group. Then we predict the task duration the control group would need to achieve the same accuracy as the Copilot group. Finally, we compute the difference between task durations for the two groups, using a bootstrap to compute the level of certainty of this finding.

Holding accuracy/quality constant, the time savings by task are as follows:

	% savings for Copilot users	P-value	Findings
Overall	23.1%	0.00	Copilot users did the **overall task 23.1% faster**. The difference is statistically significant.
Script analysis	10.1%	0.15	Copilot users did the **script analysis task 10.1% faster**.
Incident report	20.5%	0.01	Copilot users did the **incident report task 20.5% faster**. The difference is statistically significant.
Incident summary	46.2%	0.00	Copilot users did the **incident summary task 46.2% faster**. The difference is statistically significant.
Response task	-26.3%	0.00	Copilot users did the **response task 26.3% slower**. The difference is statistically significant.

We note also that Copilot currently often takes 20+ seconds to open. This necessarily slowed the Copilot users. Product improvements should reduce this duration and further increase the time savings for users with Copilot. This might account for why the response task is the only one that took Copilot users longer to complete, as the control group users spent by far the least time on this task (61% less time than the next shortest task).

We used statistical methods to hold accuracy/quality constant because there is a large difference in accuracy/quality, as discussed in the prior section. It is uninformative to compare speeds across groups when one group is systematically more accurate than another. Hence, our examination is of the time that would be required to achieve comparable accuracy.

Sentiment

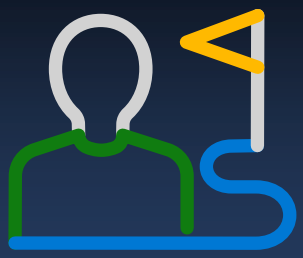
Users who had access to Copilot rated it favorably. We presented them with the following statements, with sliders to range from complete disagreement (scored as 0) to complete agreement (scored as 100). They were **all above 80**, as shown in the first numeric column below.
More than 80% of users indicated general agreement (slider positions strictly greater than 50), as shown in the second numeric column.

97%
want Copilot again next time

	Average agreement	Proportion agreeing
Copilot reduced my effort on this task	81.8	84.7%
Copilot made me more productive	87.1	91.8%
Copilot helped me improve the quality of my work	86.3	93.2%
I would want to have Copilot the next time I do this task	89.9	97.2%

Findings - novices

The findings in this section are based on our study of 149 novices.



Accuracy and quality[3]

Overall, across the three tasks with multiple choice questions, Copilot subjects got

35% **more questions correct ****

Subjects with Copilot were **significantly more accurate in a range of tasks**. Using Script Analyzer, Copilot subjects were 34%** more accurate in answering questions about the various scripts used by the attacker. Using incident report, Copilot subjects were 25%** more accurate in answering questions about the incident facts. Using guided response, Copilot subjects were 43%** more accurate in answering questions about the appropriate remediation steps.

These are high numbers. We think two factors would cause Copilot to be somewhat less extraordinary in performance. One, our tasks are closely linked to Copilot's capabilities. We show that Copilot is great at helping analysts figure something out (e.g. analyze what a given script does). Is Copilot as good at "figuring out what needs figuring out"? We have some evidence that it does – the incident report task is cross-cutting. But in the real world, tasks will not align as closely with Copilot capabilities.

Two, our analysts are security rookies with minimal skills. Experienced security analysts have a higher baseline, which leaves less room to improve.

[3] Findings in this section are revised somewhat from the draft circulated in November 2023. We adjusted our data cleaning criteria (including as discussed in the next footnote) and improved the questions used for LLM grading of essays. Results are qualitatively and directionally unchanged.

Findings as to increases in accuracy:

	Overall	Copilot	Control	P-val
Overall	8.52	9.67	7.15	0.00
Script analysis	3.40	3.84	2.87	0.00
Incident report	3.24	3.59	2.88	0.00
Response	3.39	3.87	2.70	0.00

Findings

Copilot users were **35% more accurate** at the overall task. The difference is statistically significant.

Copilot users were **34% more accurate** at the script analysis task. The difference is statistically significant.

Copilot users were **25% more accurate** at the incident report task. The difference is marginally statistically significant.

Copilot users were **43% more accurate** at the response task. The difference is statistically significant.

Just as with the security professionals, we asked subjects to write incident summaries based on the findings in Defender XDR. Copilot subjects were free to copy-and-paste the incident summary from the Copilot incident summary skill or rewrite as they saw fit. We used the same LLM grading approach to determine whether essays contained the same 15 key facts our security experts identified. The LLM found that essays from **Copilot users had an average of 10.6 of these facts, versus just 5.9 in essays from control users**. That's almost double** as many key facts in the Copilot-assisted essays. AI also praised the quality of writing, granting a 19% higher score on a five-point scale.

	Overall	Copilot	Control	P-val
Content	8.33	10.61	5.89	0.00
Quality	3.14	3.40	2.86	0.00

Findings

Copilot users got an **80% higher content score** on their essay.
The difference is statistically significant.

Copilot users got a **19%** higher quality score** on their essay.
The difference is statistically significant.

Speed

Holding accuracy/quality constant, Copilot users were

26% faster**

(This analysis compares Copilot users' speed to the speed control users would have needed to achieve the same level of accuracy/quality.)

We reach this finding via a linear regression framework that proceeds in three steps. First, we estimate the task duration as a function of accuracy for the control group. Then we predict the task duration the control group would need to achieve the same accuracy as the Copilot group. Finally, we compute the difference between task durations for the two groups, using a bootstrap to compute the level of certainty of this finding.

Holding accuracy/quality constant, the time savings by task are as follows:

	% savings for Copilot users	P-value	Findings
Overall	25.9%	0.00	Copilot users did the **overall task 25.9% faster**. The difference is statistically significant.
Script analysis	22.0%	0.01	Copilot users did the **script analysis task 22.0% faster**. The difference is statistically significant.
Incident report	16.7%	0.04	Copilot users did the **incident report task 16.7% faster**. The difference is statistically significant.
Incident summary	28.5%	0.07	Copilot users did the **incident summary task 28.5% faster**. The difference is weakly statistically significant.
Response task	19.2%	0.11	Copilot users did the **response task 19.2% faster**.

We note also that Copilot **often took 20+ seconds to open**. This necessarily slowed the Copilot users. Product improvements should reduce this duration and further increase the time savings for users with Copilot.

As with professionals in the previous section, we used statistical methods to hold accuracy/quality constant because there is a large difference in accuracy/quality. In fact, it seems many control users gave up and guessed or left questions blank—which they can do very quickly, but which doesn't give a realistic sense of how long they would require to make a good-faith effort. Hence, our examination considers the time that would be required to achieve comparable accuracy.

Sentiment

Users who had access to Copilot rated it favorably. We presented them with the following statements, with sliders to range from complete disagreement (scored as 0) to complete agreement (100). They were **all above 80**, as shown in the first numeric column below. More than 90% of users indicated general agreement (slider positions strictly greater than 50), as shown in the second numeric column.

93%

want Copilot again next time

	Average agreement	Proportion agreeing
Copilot reduced my effort on this task	83.2	90.0%
Copilot made me more productive	86.0	90.4%
Copilot helped me improve the quality of my work	85.8	91.8%
I would want to have Copilot the next time I do this task	89.8	93.2%

We also asked both treatment and control users their agreement with standard statements about the task. The Copilot users were all more favorable than control users, **and 4 of 9 statements had statistically significant differences**, as indicated below.

	Overall	Copilot	Control	P-val	% diff[4]
I felt effective doing this task	69.1	73.5	64.3	0.04	14%
I felt productive doing this task	76.1	79.8	72.1	0.02	11%
This task was draining	44.5	43.5	45.6	0.47	5%
This task was a lot of effort	58.7	53.5	64.0	0.02	16%
I would like a job like this as my full-time job	73.0	78.2	67.5	0.09	16%
I felt in control while doing this task	65.1	70.1	59.9	0.07	17%
I felt secure while doing this task	75.1	76.4	73.8	0.60	4%
I felt inadequate while doing this task	41.8	32.1	52.0	0.00	38%
I felt uncertain while doing this task	45.4	41.6	48.7	0.01	15%

Users could feel how much time Copilot saved them. When asked how much time Copilot saved, users with Copilot said it saved 38 minutes on average. In fact, control users finished just five minutes slower (at least in part because many of them gave up, as discussed above). Perhaps Copilot users are trying to answer *how much longer it would take to do the task if I tried to achieve the same accuracy*, rather than just "how much longer did the control users take." But the fact is, users with Copilot **sharply overestimated the time savings, by approximately 7x**, **consistent with them enjoying the tool** and being glad they had access to it.

Our research subjects were rookies. They recognized the difficulty of the task and the benefit that Copilot provided.

[4] All differences were in the direction of positive sentiment towards Copilot. Note that some of the statements are written in the positive and others in the negative.

Comparison of professionals and novices

The major difference between security professionals and novices is that the accuracy gains from Copilot are much greater for novices. We see this finding as unsurprising. For one, novices have no specialized training in this area, whereas the professionals are trained, skilled experts. Although Copilot confers smaller accuracy gains for security professionals, it allows them to perform the tasks faster without sacrificing accuracy, and professionals still do get a statistically significant increase in overall accuracy.

Holding accuracy constant, **professionals and novices got similar results**. For professionals, overall time savings was 23.1%**.
For novices, the Copilot users did the overall task 25.9%** faster, holding accuracy constant.

For subtasks, the **findings were qualitatively similar**, never more than five percentage points different. The notable exception is response, where novices with Copilot were 19.2%** faster than novices without, whereas professionals got slower with Copilot. As discussed above, this slow-down for professionals on the response task is likely due to the combination of the task being the shortest duration and the time it takes Copilot to load.

For sentiment, we saw **somewhat more favorable statements from novices**, but the differences never exceeded two points of average agreement or six percentage points of proportion agreeing. The greater treatment effects for novices in accuracy at least partially account for the differences in sentiment: subjects' sentiments about Copilot should reflect how much it helped them. Another possible explanation is that security professionals routinely use dozens of tools in their workflows. In this experiment, we gave them none of their standard tools. They only had Defender XDR and a new tool, Copilot. This means the novices and professionals likely have different baselines when thinking about the benefits of Copilot: novices were happy to get any help from any tool,

whereas professionals were likely to think about how much more comfortable they are with their preferred tools. In this light, the reported sentiment from security professionals might actually be seen as surprisingly favorable.

These findings broadly validate Security Copilot: it makes both novices and professionals more productive, and our test subjects can feel how much more productive they are when using this tool.

Related work

We join a literature in which a randomized controlled trial measures the causal impact of AI tools. In this line of papers, some users are randomly granted AI tools, while others use standard tools. Then the difference between these groups indicates the effect of the tools, with randomization eliminating bias from endogeneity. Representative papers in this field include [Peng, et al. 2023], [Brynjolfsson, et al. 2023] and [Noy and Zhang, 2023], which highlight how AI tools reduce task completion time and increase output quality, bringing a substantial improvement in workplace efficiency. [Choi and Schwarcz, et al. 2023], [Mollick, et al. 2023] and [Horton, et al. 2023] provide further evidence of AI's profound impact on performance of workers, students, and jobseekers, respectively.

A second line of papers finds generative AI effective at helping novices accomplish tasks traditionally performed by subject-matter experts. For example, [Brynjolfsson, et al. 2023] finds that an AI-based conversational assistant provides an average 14% productivity increase improvement, and the benefit is largest for novice and low-skilled workers. Similarly, [Dell'Acqua, et al. 2023] shows that while consultants across the skills distribution benefited from AI augmentation, the benefit was larger among those previously in the bottom half of the skills distribution (who got a 43% increase, compared to a 17% increase for those in the top half). Researchers also find that using AI improves the performance of novices in tasks such as organizational decision-making [Spitzer, et al. 2022], and data work [Sun, et al. 2022].

Generative AI can boost productivity of highly skilled workers, as shown in sectors including software development [Kalliamvakou, 2022], economic research [Korinek 2023], encoding medical knowledge [Singhal, et al. 2023], ideation and creative work [Dell'Acqua, et al. 2023], and in managerial professions [Sowa, et al. 2021]. These papers find that humans and generative AI can work together to produce more than the sum of their parts. For example, generative AI can spark the brainstorming process and allow humans to improve on a draft from an AI [Noy and Zhang, 2023].

Our study builds on research about automation of security tasks. The literature on this topic finds that risk mitigation is essential with the growth of security breaches, especially because so many security vulnerabilities operate at the gap between how systems are supposed to operate and how they actually operate [Morgan, et al. 2022]. Although many aspects of cybersecurity currently rely on human subject matter experts [Costa and Yu, 2018], researchers point out the possibility of automating error-prone and time-consuming security work. Machine learning techniques show particular promise in intelligently analyzing cybersecurity data [Sarker 2022]. Natural language processing, knowledge representation and reasoning, and rule-based expert systems modeling can also support AI-driven cybersecurity [Sarker, Hasan, et al. 2021].

Finally, by providing non-expert users with generative AI assistance, we study a novel way to increase their functional expertise in cybersecurity. Here, we join a separate literature (along with a robust commercial ecosystem that compares expert and non-expert computer users' security knowledge and ability to mitigate security risks. [Camp, et al. 2008] finds that security experts and non-experts have quite different mental models in performing security-related tasks, a divergence that calls into question whether non-experts can work productively in this field. [De Luca, et al. 2016] conduct an experiment on secure IM messaging with IT security experts and non-experts, finding that the expert view differs in its focus on technical and security properties thanks to a mental model that is more thorough and technology-focused. Furthermore, non-experts can be confused about using secure security practices due to lacking prerequisite knowledge common among experts. [Doswell 2008] finds that novice security users with a user-friendly security tool are able to understand basic security functions and mitigate possible security risks. Others posit techniques to capture knowledge from one domain expert and transfer it to another [Kline, 2023], a technique which is natural yet predictably limited.

References

Brynjolfsson, Eric and Li, Danielle and Raymond, Lindsey R. Generative AI at Work (November 2023), Available at ARVIX: https://arxiv.org/abs/2304.11771

Camp, L. and Asgharpour, Farzaneh and Liu, Debin. Experimental Evaluations of Expert and Non-expert Computer Users' Mental Models of Security Risks (2008). Available at: https://www.researchgate.net/publication/228671400_Experimental_Evaluations_of_Expert_and_Non-expert_Computer_Users%27_Mental_Models_of_Security_Risks

Choi, Jonathan H. and Schwarcz, Daniel. AI Assistance in Legal Analysis: An Empirical Study (August 13, 2023). Minnesota Legal Studies Research Paper No. 23-22.

De Luca, Alexander and Das, Sauvik and Ortlieb, Martin and Ion, Iulia and Laurie, Ben. Expert and Non-Expert Attitudes towards Secure Instant Messaging (June 2016), Twelfth Symposium on Usable Privacy and Security (SOUPS 2016), Available at USENIX Association: https://www.usenix.org/conference/soups2016/technical-sessions/presentation/deluca

Dell'Acqua, Fabrizio and McFowland, Edward and Mollick, Ethan R. and Lifshitz-Assaf, Hila and Kellogg, Katherine and Rajendran, Saran and Krayer, Lisa and Candelon, François and Lakhani, Karim R. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality (September 15, 2023). Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24-013, Available at SSRN: https://ssrn.com/abstract=4573321 or http://dx.doi.org/10.2139/ssrn.4573321

Kline, E., Bartlett, G., Lawler, G., Story, R., Elkins, M. Capturing Domain Knowledge Through Extensible Components. In: Gao, H., Yin, Y., Yang, X., Miao, H. (eds) Testbeds and Research Infrastructures for the Development of Networks and Communities (2019). TridentCom 2018. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 270. Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-12971-2_9

Korinek, Anton. Language Models and Cognitive Automation for Economic Research (2023). National Bureau of Economic Research. Working Paper Series 30957, Available at NBER: http://www.nber.org/papers/w30957

Mollick, Ethan R. and Mollick, Lilach. Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts (2023), Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4391243

Morgan, Philip L., and Collins, Emily I.M., and Spiliotopoulos, Tasos, and Greeno David J., and Jones, Dylan M. Reducing risk to security and privacy in the selection of trigger-action rules: Implicit vs. explicit priming for domestic smart devices, International Journal of Human-Computer Studies, Volume 168, 2022, 102902, ISSN 1071-5819, Available at: https://doi.org/10.1016/j.ijhcs.2022.102902.

Noy, Shakked and Zhang, Whitney. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence (March 1, 2023). Available at SSRN: https://ssrn.com/abstract=4375283 or http://dx.doi.org/10.2139/ssrn.4375283

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot (2023). Available at ARVIX: https://arxiv.org/abs/2302.06590

Sarker, I.H.. Machine Learning for Intelligent Data Analysis and Automation in Cybersecurity: Current and Future Prospects. Ann. Data. Sci. 10, 1473–1498 (2023). Available at: https://doi.org/10.1007/s40745-022-00444-2

Schwab, S. and Kline, E.."Cybersecurity Experimentation at Program Scale: Guidelines and Principles for Future Testbeds," 2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Stockholm, Sweden, 2019, pp. 94-102, doi: 10.1109/EuroSPW.2019.00017.

Singhal, K., Azizi, S., Tu, T. et al. Large language models encode clinical knowledge. Nature 620, 172–180 (2023). Available at: https://doi.org/10.1038/s41586-023-06291-2

Sowa, Konrad and Przegalinska, Aleksandra and Ciechanowski, Leon. Cobots in knowledge work: Human – AI collaboration in managerial professions, Journal of Business Research, Volume 125, 2021, Pages 135-142, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2020.11.038. Available at: https://www.sciencedirect.com/science/article/pii/S014829632030792X

Spitzer, Philipp and Kühl, Niklas and Goutier, Marc. Training Novices: The Role of Human-AI Collaboration and Knowledge Transfer (July 1, 2022), Available at ARVIX: https://arxiv.org/abs/2207.00497

Sun, L., Liu, Y., Joseph, G., Yu, Z., Zhu, H., & Dow, S. P.. Comparing Experts and Novices for AI Data Work: Insights on Allocating Human Intelligence to Design a Conversational Agent (2022). Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 10(1), 195-206. Available at: https://doi.org/10.1609/hcomp.v10i1.21999

Van Inwegen, E., Munyikwa, Z. and Horton, J., Algorithmic Writing Assistance on Jobseekers' Resumes Increases Hires (October 2023), Available at ARVIX: https://arxiv.org/abs/2301.08083