Generative AI in Real-World Workplaces

The Second Microsoft Report on AI and Productivity Research

Editors:
Sonia Jaffe, Neha Parikh Shah, Jenna Butler, Alex Farach, Alexia Cambon, Brent Hecht, Michael Schwarz, and Jaime Teevan

Contributing Researchers:
Reid Andersen, Margarita Bermejo-Cano, James Bono, Georg Buscher, Chacha Chen, Steven Clarke, Scott Counts, Eleanor Dillon, Ben Edelman, Ulrike Gruber-Gremlich, Cory Hilke, Ben Hanrahan, Sandra Ho, Brian Houck, Mansi Khemka, Viktor Kewenig, Madeline Kleiner, Eric Knudsen, Sathish Manivannan, Max Meijer, Jennifer Neville, Nam Ngo, Donald Ngwe, Ried Peckham, Sida Peng, Nora Presson, Nagu Rangan, Reetchatha Rangareddy, Sean Rintel, Roberto Rodriguez, Katie Rotella, Tara Safavi, Advait Sarkar, Ava Elizabeth Scott, Abigail Sellen, Chirag Shah, Auste Simkute, Tyler Smith, Shwetha Srinath, Siddharth Suri, An-Jen Tai, Lev Tankelevitch, Mengting Wan, Leijie Wang, Ryen White, and Longqi Yang
(with additional support from the entire AI and Productivity team at Microsoft)

ABSTRACT

This report presents the most recent findings of Microsoft's research initiative on AI and Productivity, which seeks to measure and understand the productivity gains associated with LLM-powered productivity tools like Microsoft Copilot. The report synthesizes research results from over a dozen recent studies conducted by researchers at Microsoft, with a focus on studies of generative AI in actual workplace environments. One of these is, to our knowledge, the largest, randomized controlled trial of the introduction of generative AI into organizations. Overall, the research suggests that generative AI is already aiding workers in becoming more productive in their day-to-day jobs in significant ways. However, the influence of generative AI is subject to variation by role, function, and organization and is contingent upon adoption and utilization. The report explores these variations and underscores the potential for AI to have even greater impact as individuals and organizations recalibrate their work practices to harness AI in the places where it provides the most value.

Please cite this report as:

Jaffe, S., Shah, N.P., Butler, J., Farach, A., Cambon, A., Hecht, B., Schwarz, M. and Teevan, J. eds. 2024. Generative AI in Real-World Workplaces: The Second Microsoft Report on AI and Productivity Research. Microsoft.

1 INTRODUCTION

There is tremendous interest in how AI can increase people's productivity at work. To help meet this interest, in December 2023, Microsoft released a first *AI and Productivity Report* (Cambon et al. 2023) synthesizing the results of many Microsoft studies on AI and productivity. These studies contributed to a large and growing literature from around the world and a wide variety of disciplines. Although there are exceptions, this literature largely points to a broad conclusion: Generative AI tools have the potential to introduce a substantial step-function increase in productivity for tasks performed by information workers (e.g., Noy and Zhang 2023; Dell'Acqua et al. 2023; Brynjolfsson et al. 2023; Peng et al. 2023).

However, much of this existing literature on AI and productivity is limited in that it consists primarily of lab-based studies. In these studies, participants used generative AI tools to complete researcher-designed tasks in a controlled, simulated work environment, largely with a focus on tasks that the researchers hypothesized would be amenable to generative AI. Now that a much larger population of workers has access to generative AI tools, we can begin to understand the impact of these tools outside of a lab setting, as people perform their everyday jobs. This has begun to allow researchers at Microsoft and elsewhere to study how the first wave of generative AI tools impacts information work in real-world contexts.

Accordingly, this second Microsoft AI and Productivity Report focuses on Microsoft studies that explore how people apply Copilot and other generative AI tools to their regular work. The report also describes learnings from a small set of additional lab experiments that suggest new ways that we might see the impact of Copilot in-the-wild in future studies. Overall, the results – including those from what we believe is the single largest randomized controlled trial on the introduction of generative AI in real workplaces – point to several high-level observations:

- Generative AI is already helping people be measurably more productive in their day-to-day jobs.
- As expected, the productivity story in real-world workflows is more complex than observed in lab studies.
- Productivity gains associated with generative AI, including time and accuracy, vary by role, function and organization.
- Variance in adoption and utilization influences AI's impact.
- Early studies suggest generative AI may affect the cognitive effort required for task completion.

The goal of this report is to synthesize learnings from studies from around the company versus to completely describe each individual study. Many of the studies are or will be the subject of dedicated reports or research papers, and we have provided links to those documents where they are already available. Most of the studies have not yet been through peer review and as such they have not had the chance thus far to incorporate external reviewer feedback. Further, before continuing, it is important to acknowledge that all work here was funded by Microsoft, which has a commercial interest in improving and demonstrating the degree to which Copilot increases worker productivity.

2 RELATED WORK

Researchers outside of Microsoft also have been moving to study generative AI's impact on productivity in real-world contexts. This section highlights a few of the most notable studies in this space. These studies consistently show that the gains predicted by lab studies do indeed translate into significant impact when AI is used for real work. Further, they begin to reveal some nuances in how AI is used, highlighting the importance of contextual factors such as skill or task selection on AI's impact and providing early evidence that the presence of AI may impact people's behavior and the larger ecosystem.

Brynjolfsson et al. (2023) introduced one of the earliest studies of generative AI in real-world work environments. They studied an AI-based conversational assistant for customer service agents in a call center, and found that agents with the assistant resolved 14% more issues per hour than those without the assistant. Consistent with what has been observed in some lab studies (e.g., Noy and Zhang 2023), the largest impact was on novice and low-skilled workers, with very little effect on experienced or highly-skilled workers.

However, observing larger benefits for less-skilled workers is certainly not universal. Otis et al. (2024) examined the effects of a generative AI-powered entrepreneurship support tool on Kenyan entrepreneurs' business performance via an index measure based on profit and revenue. They found that entrepreneurs with above-median performance prior to the start of the experiment saw gains of 0.19 standard deviations in performance when using the AI tool, while entrepreneurs with below-median performance saw a decrease of 0.09 standard deviations. Though the two groups used the tool similar amounts, they tended to ask different types of questions. This finding emphasizes the importance of contextual factors in the productivity gains seen by using generative AI, a key observation of this report as well.

In addition, some early real-world studies show that the presence of AI may have cascading effects, not just affecting productivity on a given task, but also changing which tasks people choose to do. For example, Wiles and Horton (2024) explored how having an LLM generate a first draft of a job posting affected postings and hiring on a large online labor market. They found that the AI tool decreased time spent writing posts and increased the number of posts completed but had no effect on the number of hires. The researchers suggest this may be because the additional jobs that get posted were less important than their other jobs and using AI to draft may have caused employers to exert less effort in writing the job posts, leaving them with fewer well-matched applicants. In another example, Yeverechyahu et al. (2024) studied generative AI effects on coding activity, both quantity and type. Specifically, the authors compared open-source repositories for packages in Python (which was supported by GitHub Copilot) to R (which was not). They found a significant jump in contributions due to GitHub Copilot. Of note, the increase was larger for contributions categorized as "maintenance solutions" than for those categorized as "new code development," which require more extrapolative thinking.

In addition to impacting human behavior, the real-world presence of generative AI may also influence the future behavior of AI systems themselves. Rio-Chanona et al. (2023) provide evidence that the availability of generative AI programming tools substantially reduced participation in online programming forums that produce important training data for these tools. This suggests that in some contexts a "paradox of reuse" dynamics (Taraborelli 2015; McMahon et al. 2017) might be emerging in the generative AI ecosystem. These dynamics could significantly harm productivity gains if not properly addressed (Vincent 2022).

3 STUDIES AND RESULTS

We now provide an overview of studies recently conducted by researchers at Microsoft, focusing in particular on those that speak to real-world implications of generative AI.

Studies of workers using AI on the job

Early Access Program Telemetry Study (*Eleanor Dillon, Sonia Jaffe, Sida Peng, and Alexia Cambon*)

Working with over 60 organizations and including over 6000 individual employees across a wide range of industries and occupations, researchers conducted a large-scale randomized controlled field experiment of Copilot for Microsoft 365 – relying on participants using the tool in their day-to-day work as opposed to on researcher-defined tasks in a lab context. We believe this research is the largest controlled study of productivity impacts in real-world generative AI deployments to date.

Researchers worked with organizations in the Copilot for Microsoft 365 (M365) Early Access Program to create a randomized control trial. Copilot for Microsoft 365 combines generative AI tools in applications such as Word, Excel, PowerPoint, Outlook, Teams, and others. Each organization set aside at least 50 licenses to be randomly assigned among 100 or more Microsoft 365 users nominated by the organization. Researchers partnered closely with IT administrators and business decision-makers in each of the over 60 participating organizations to explain the need for randomization and obtain buy-in. To ensure privacy, researchers

looked only at aggregate effects and did not analyze or report individual-level data.

Using metadata from Microsoft 365 in these organizations, researchers compared how email, meeting, and document behavior differed based on being assigned a Copilot license. Researchers found that on average, those with Copilot for Microsoft 365 read 11% fewer individual emails and spent 4% less time interacting with them, compared to people without Copilot. Some organizations saw larger effects with relative decreases of up to 20 or 25% in both emails read and time spent interacting with email. The top graph in Figure 1 shows the effects across organizations. The researchers hypothesize that the summarize emails feature in Copilot for Outlook and the Copilot chat function may have allowed workers to retrieve information without reading or rereading individual emails.

The effects of Copilot on the number of meetings attended (via Microsoft Teams) were more complex, with some organizations seeing significant increases, others seeing significant decreases, and others seeing no significant effect. Of the 47 organizations that have been in the study the longest, 10 saw a statistically significant decrease in attended meetings, with an average decrease of .39 meetings per day on a pre-Copilot average of 3 meetings. On the other hand, 14 customers saw a statistically significant increase in meetings with an average increase of .36 meetings per day on a pre-Copilot average of 2 meetings. The difference in baseline average suggests that increases in meetings attended were more likely at organizations with low levels of baseline Teams usage. The remaining customers did not see a statistically significant change in the number of Teams meetings. The middle graph in Figure 1 plots the effect for each customer.

Teams Copilot provides meeting summaries and uses the transcript to answer questions users may ask of it during or after a meeting, but only works for meetings that are executed in Teams. Thus, having access gives people an additional reason to have meetings in Teams (instead of only in-person or via another app), so increases could reflect increases in the use of Teams that are not indicative of increases in overall meetings. More generally, if Copilot makes meetings both more effective and more efficient, that can generate conflicting effects: more efficient meetings require less time and fewer follow-up meetings, but if meetings become more effective collaboration tools, they may be used for a wider range of projects or tasks.

With respect to documents, people with Copilot also created and edited more documents than those without Copilot. Overall people edited 10% more documents, with heavy users of Word, Excel, and PowerPoint seeing an increase of 13% (on a higher baseline). Some organizations saw increases in the 25-30% range. One hypothesis is this is an early sign of the writing and creation assistance that Copilot provides making it easier to produce and revise output. Alternatively, people may be using some of the time they save with Copilot to do additional document creation and editing.

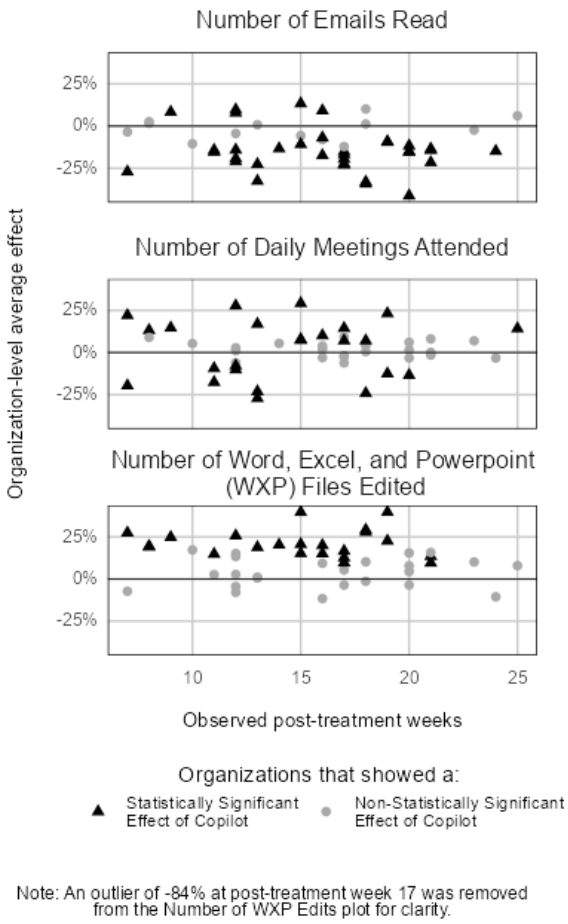

Figure 1. Effects across organizations of access to Copilot for M365 on emails read, scheduled meetings attended, and files edited in Word, Excel, and PowerPoint.

This study is still under way, and researchers are planning to explore additional outcomes (e.g., amount of time spent per document) as well as spillovers and team effects (e.g., the impact of a worker's collaborators having Copilot). This study is limited by its focus on work processes. That is, though telemetry can provide an objective measure of activity, there is not a direct mapping between the observed outcomes (number of documents, emails, etc) and productivity, performance or business outcomes. Moreover, to preserve privacy, the study observes activity, not the content created, so it cannot study quality or how well output aligns with people's goals or intents.

Work Trend Index Survey

More study details available in AI at Work Is Here. Now Comes the Hard Part (Microsoft and LinkedIn 2024)

To understand the impact of generative AI on workplace productivity and satisfaction, Microsoft conducted the 2024 Work Trend Index Survey. This 20-minute, anonymous survey was administered by Edelman Data & Intelligence to 31,000 full-time employed or self-employed knowledge workers across 31 countries

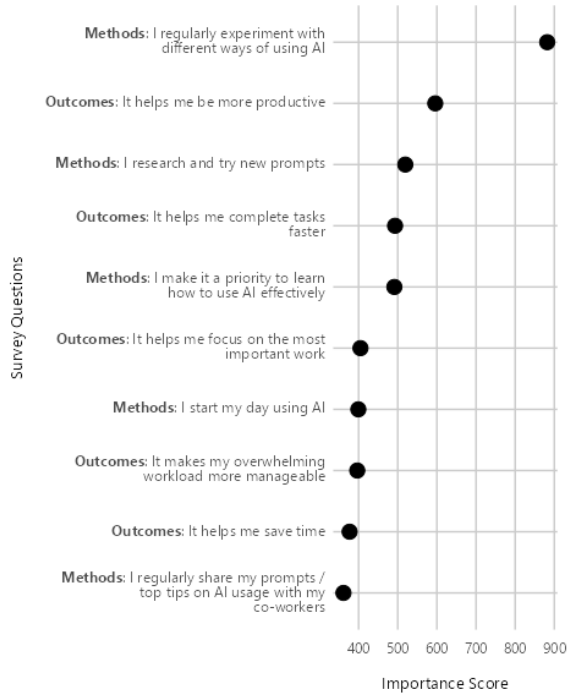

Figure 2. Variable importance in predicting AI power usage

between February 15, 2024, and March 28, 2024. The survey aimed to capture user sentiments and experiences with generative AI broadly as opposed to focusing on any specific generative AI tool such as Copilot.

One key finding from the survey is the widespread use of unsanctioned AI tools among employees. The survey revealed that, of respondents who used AI, 78% used at least some AI tools not provided by their organization. This highlights a significant phenomenon where many employees turn to external AI resources to meet their needs.

Additionally, a significant focus of the Work Trend Index data analyses is "AI Power Users," which researchers defined as individuals reporting being familiar with generative AI, using it at work at least several times a week, and saving more than 30 minutes a day by using it. Overall, 29% of respondents who used AI fell into this bucket. Power users had noticeably lower use of unsanctioned AI (66% vs. the non-power user average of 83%, $p<.05$).

Researchers sought to understand which factors from the survey were most predictive of the power user classification; they focused on several survey questions categorized into three areas:

- Actions: Actions related to generative AI at work.
- Methods: Methods of AI usage.
- Outcomes: Feelings or outcomes related to respondent AI usage.

The survey responses were analyzed to build a model identifying the key predictors of AI power user classification. The data preparation and modeling process included addressing class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) and evaluating model performance through cross-validation. Researchers implemented two predictive models: Random Forest and Logistic Regression. The Random Forest model outperformed Logistic Regression with an accuracy of 0.744 and a ROC-AUC score of 0.737, compared to Logistic Regression's accuracy of 0.657 and ROC-AUC score of 0.695. Consequently, the Random Forest model was trained on the entire dataset to pinpoint the key predictors of AI power usage.

As seen in Figure 2, regular experimentation with AI emerged as the most significant predictor of AI power usage classification. This factor was a stronger predictor of power user classification than other AI specific methods, actions, or outcomes. The importance score, measured by a Random Forest statistical model, should be interpreted relatively, as it shows how much each feature helps in predicting AI power usage compared to others. Higher scores indicate greater importance. In this analysis, scores range from 361 to 882, highlighting the significant factors influencing AI power user classification within this dataset and model.

As with all surveys of this type, it is important to view all the above results through the lens of the limitations of the methodology. While the analysis reveals significant associations, causation cannot be conclusively established due to the observational nature of the data. Similarly, self-selection bias, response bias, and unmeasured confounding variables such as workplace culture and managerial support could influence the outcomes.

Copilot Usage in the Workplace Survey (*Alexia Cambon, Alex Farach, Margarita Bermejo-Cano, and Eric Knudsen*)

More study findings available in AI Data Drop: The 11 by 11 Tipping Point

While the Work Trend Index survey described above focused on generative AI in general, researchers also conducted a broad survey focused specifically on Copilot for Microsoft 365, asking enterprise Copilot users about their perceived benefits, time savings, and overall job satisfaction. This 20-minute, anonymous survey, which is ongoing, is being distributed to people with Copilot licenses at participating customer organizations from October 1, 2023, to November 1, 2024. Analysis here is based on 885 responses collected up to February 1, 2024, from people who had used Copilot for more than three weeks at the time of survey response.

The survey results suggest that people who used Copilot for an extended period receive significant benefits from doing so. Researchers analyzed three distinct categories of usage durations: 3-6 weeks, 7-10 weeks, and more than 10 weeks. The analysis employed a 5-point Likert scale, where 1 represents "Strongly Disagree" and 5 represents "Strongly Agree."

Respondents who had been using Copilot for more than 10 weeks reported greater benefits compared to those with shorter usage durations. For example, for the question "Using Copilot in Teams allows me to attend fewer meetings," those using Copilot for 3-6

weeks had an average response of 2.66 (variance 0.81) and those with more than 10 weeks of usage had an average of 3.06 (variance 1.34). For the question "Using Copilot helps me to enjoy my work more" the average for the 3-6 week group was 3.4 and for the over 10-weeks group it was 3.6. All results reported in this analysis are statistically significant, as confirmed by ANOVA tests ($p < 0.05$).

It is essential to acknowledge the limitations of this self-reported data. While researchers found significant associations, establishing causation is challenging. People who are more inclined to use productivity tools like Copilot may also be those who naturally experience higher job satisfaction, creating potential self-selection bias. Additionally, unmeasured factors such as managerial support or workplace culture could influence both Copilot usage and job satisfaction. This study also relied on self-reported data that introduces the possibility of response bias.

Study on Generative Search Engines and Task Complexity (*Siddharth Suri, Scott Counts, Leijie Wang, Chacha Chen, Mengting Wan, Tara Safavi, Jennifer Neville, Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, Sathish Manivannan, Nagu Rangan, and Longqi Yang*)

More study details available in The Use of Generative Search Engines for Knowledge Work and Complex Tasks (Suri et al. 2024)

Search is a common task in real-world workflows. To understand how the use of AI-augmented search differs from traditional search, researchers analyzed 80,000 randomly selected, de-identified conversations from the consumer version of Copilot in Bing and traditional Bing searches. They used GPT-4 to classify each conversation and search by topic domain. They found that chats with Bing Copilot tend to focus on topics related to knowledge work, such as "Translation and language learning," "Creative writing and editing," and "Programming and scripting." Overall, 72.9% of the Copilot conversations are in knowledge work domains compared to 37% of Bing Search sessions. The researchers also used GPT-4 to directly classify whether the task associated with each Copilot conversation and or search session was knowledge work (instead of classifying based on the category) and see a similar pattern.

Researchers then used GPT-4 to classify the main task associated with each conversation or search sessions according to Anderson and Krathwohl's Taxonomy (Anderson and Krathwohl 2001), which defines six categories from lowest complexity (for a human) to highest: Remember, Understand, Apply, Analyze, Evaluate, and Create. Over three-quarters of traditional search sessions, but less than half of Copilot conversations were for "Remember" tasks. Grouping "Remember" and "Understand" as low-complexity tasks and the rest as high-complexity, the authors found 13.4 % of traditional search sessions and 37% of Copilot sessions were high-complexity. That is, AI-augmented search tended to be in higher complexity domains than traditional search.

Researchers interpret the shift in domain and complexity of tasks between traditional search and Copilot as generative AI helping people with tasks that used to be done with much more human effort; LLMs shift the frontier of which tasks machines can help with – and how helpful they are. Researchers caveat that these results are based on early usage of Bing Copilot and patterns may change as the tools develop and users gain experience working with them. Nonetheless, the study suggests that LLMs will affect substantial changes in how people use search-based tools and accomplish knowledge work tasks more broadly.

The study could not identify when people were using consumer Copilot for their jobs, but the high number of knowledge work tasks is consistent with the finding in the Work Trend Index report that many employees were using generative AI tools not provided by their companies in their work.

Specific Roles and Functions

In addition to the studies above, which span across a range of knowledge workers, some studies also look at how results differ across roles and study generative AI tools developed for use cases of a specific role or profession. Following a comparison across roles, we take a closer look at the software development function. Because of the earlier availability of GitHub Copilot, there is a lot of research in that area which may lend insight to effects to be expected in other types of information work.

Comparing across Roles in Copilot Usage in the Workplace Survey *(Alexia Cambon, Alex Farach, Margarita Bermejo-Cano, and Eric Knudsen)*

The Copilot Usage in the Workplace Survey described above helped researchers understand broad patterns, but also allowed us to look at data by job type to see the impact on specific roles and functions, focusing on two dimensions: adoption and perceived benefits. All reported differences are statistically significant at $p<.05$, using ANOVA tests, with the Benjamini-Hochberg False Discovery Rate (FDR) control procedure applied.

When asked about usage (1=never, 5=daily), respondents across nearly all functions reported using Copilot in Teams at least weekly, with some functions like sales and product development reporting daily usage (average response scores of 4.66 and 4.55, respectively). Others functions like legal and supply chain reported somewhat lower usage (4.03 and 3.88, respectively). Reported usage of Copilot in Outlook was generally slightly lower, but with similar patterns across roles.

Respondents with communication-focused responsibilities involving repetitive and/or content creation tasks supported by current AI capabilities also reported the most benefits from Copilot, including productivity, fulfillment, work quality improvements, and efficiency. In contrast, those in roles involving more variable and/or complex tasks not yet fully optimized by current AI capabilities, including legal and R&D, reported fewer benefits. Some distinctions also may be attributed to highly regulated industries or sensitive use cases. It is likely, however, that AI tools will be improved over time to support these scenarios.

Specifically, when asked about increased productivity with Copilot ("When using Copilot I am more productive" with 1="Strongly Disagree" and 5="Strongly Agree"), the reported effect was highest among professionals in customer service and sales (mean of 4.2 and 3.97, respectively), and lowest among legal professionals (mean of 3.0). For fulfillment ("When using Copilot I feel more fulfilled in my work") customer service and sales functions reported the highest mean agreement (3.53 and 3.41, respectively) with R&D and legal functions experiencing the lowest mean agreement (2.74 and 2.90, respectively). Additionally, customer service and product development professionals rated the improvement in work quality ("Using Copilot helps improve the quality of my work") highest, with mean scores of 4.20 and 3.93, respectively. In contrast, legal professionals again reported lower improvement scores, averaging 3.19.

In terms of efficiency and information management, customer service, creative, and sales professionals reported significant ease in catching up on missed meetings and retrieving necessary information, with mean response scores of 4.27, 4.40, and 4.45, respectively. Again, legal and operations functions had the lowest mean scores for meeting efficiency at 3.28 and 3.75, respectively.

As noted above, this study shares the limits of all surveys in relying on self-reports. Moreover, there may be differences across professions in how people perceive the subjective metrics like quality and fulfillment that the survey asked about.

Towards Effective AI Support for Developers: A Survey of Desires and Concerns (*Mansi Khemka and Brian Houck*)

More details available in Towards Effective AI Support for Developers: A Survey of Desires and Concerns (Khemka and Houck, 2024).

Microsoft researchers surveyed 800 Microsoft developers and explored the opportunities and concerns that they have with using AI in their work. Responses indicated that developers most want to see AI help with automating routine tasks, like generating unit tests and writing documentation, which they find monotonous but essential. Specifically, 44% of respondents highlighted generating tests as a top area where AI could alleviate the burden and improve developer experience. Additionally, 42% noted AI's potential in analyzing code for defects and optimizations, seeing it as a virtual pair-programming partner. Writing documentation was another area of interest, with 37% seeing AI's potential in automating this crucial but often neglected task.

However, developers also voiced significant concerns. The top worry (29%) was that AI might not be as helpful as expected. Another major concern (21%) was that AI might introduce defects or vulnerabilities, emphasizing the need for thorough validation and human oversight. Job security was a worry for 10% of respondents, reflecting fears of AI encroaching on their roles.

These learnings suggest developers view AI as helpful to improving aspects of their workflows, even as they remain uncertain of AI's promise and concerned about threats to their job security. To mitigate the negative effects of this uncertainty on productivity and innovation, and to maintain developers' trust and satisfaction, organizations may identify ways to integrate AI into developers' workflows effectively. These may include acknowledging and addressing concerns and offering training programs.

Problem-Solving Styles and Confidence Generating Prompts for GitHub Copilot (*Steven Clarke and Ben Hanrahan*)

This study explored how developers' problem-solving styles influence their confidence when generating prompts for GitHub Copilot. The authors hypothesized that variations in developers' problem-solving approaches and workstyles would significantly influence their interactions with Copilot, thereby affecting their confidence and productivity outcomes.

To explore this hypothesis, the authors employed the GenderMag survey (Burnett et al. 2016; Anderson et al. 2024), a tool specifically designed to investigate the impact of differences in people's problem-solving styles when working with technology. This survey was used along with additional questions about confidence in Copilot prompting overall and for different scenarios. A third-party recruiting firm recruited participants who worked in programing roles, had done so for at least six months, and used GitHub Copilot at work. The survey was sent to 250 people, yielding 212 usable responses. To analyze the data, researchers ran a regression model to measure the extent to which years of experience, time using Copilot, and each GenderMag trait, (Computer self-efficacy, risk-aversion, info-processing style, motivation for technology use, and learning style), explained respondents' confidence in each scenario.

The study found that the duration for which developers have been using GitHub Copilot was the most significant factor explaining their prompting confidence. They also found that confidence in prompting is inversely related to the number of years of professional software development experience. While this may seem counterintuitive, it could be because they are more familiar with or attached to existing workflows or because more experienced developers are better able to spot errors and inaccuracies in Copilot responses. If they attribute those errors to their prompts, it could make them less confident that they can create successful prompts. The analysis also showed that developers with a comprehensive approach to information processing and developers who are motivated to use technology for its own sake are also more confident in generating prompts. These findings echo those above suggesting benefits of usage increase over time.

GitHub Copilot and Engineering System Satisfaction (*An-Jen Tai, Shwetha Srinath, and Reetchatha Rangareddy*)

This study considered how engineering system satisfaction (as measured by a net satisfaction score or NSAT) changed for Microsoft employees who adopted GitHub Copilot compared to those who did not. Researchers examined anonymized data on

>30,000 software engineers, some of whom had installed and used GitHub Copilot between the two waves of Microsoft's bi-annual Employee Signals Survey, combined with their survey responses on satisfaction with the engineering systems.

The 95% confidence interval for the difference-in-differences estimate (the change for the adopters minus the change for the non-adopters) was (-2, 4.1), suggesting that Copilot did not have a significant effect on employee satisfaction with the engineering systems. In addition to not finding a statistically significant difference, the study suggests that the true difference is less than a 4pt change, which is not considered substantial since the NSAT scale can range from 0 to 200 and the average moves around a few points from survey wave to survey wave. This is perhaps unsurprising since coding is just a part of what constitutes an engineering system for most developers. For example, a prior study found developers only spend 21% of their time writing code, with the other time spent doing things like reviewing code, attending meetings, doing email, and reading technical websites (Meyer et al. 2017). Satisfaction may have been driven primarily by the other tools developers used. The lack of effect could also be because some of the users may have tried GitHub Copilot, but not used it regularly or they lacked training or manager support for use.

A Selection of New Lab Studies

While the above research focuses on the use of generative AI in the wild, we are also exploring in a lab setting some of the important trends that real-world use highlights. Given AI's impact appears to vary by role and function, several of these lab studies explore this, diving more deeply into software development and extending the analysis to other important roles like sales and security. Further, because Copilot is deployed globally, we're also starting to see variation across languages, and thus present research studies looking at AI in multilingual contexts. Finally, the complex trade-offs people are starting to make to incorporate AI into their work practices suggests the cognitive mechanisms underlying its use are important to understand, and we share some early work in that space as well.

Comparing the Effect of Different Task Types on Effective Use of GitHub Copilot (*Steven Clarke and Ben Hanrahan*)

This study investigated the conditions under which a developer might expect to benefit most from GitHub Copilot. The authors recruited 23 Java developers with at least one year of professional experience and asked them to perform one of two different tasks. Half worked on a task that involved writing new code using familiar components and concepts, and the other half worked on a different task that involved modifying existing code using unfamiliar components and concepts. For each task, half of the participants completed the task using Copilot and the other half did not use it.

All participants were allowed to use any online resources they wanted, but those in the Copilot group were first given a 10-minute overview of GitHub Copilot and encouraged to use that. Even with these small sample sizes, the researchers found evidence that the type of task matters for the impact that Copilot has on the developer. With the familiar task Copilot use resulted in 36% time-savings ($p<.05$) and 48% fewer issues ($p=.12$). In contrast, no substantial difference was observed between Copilot and non-Copilot groups for the less familiar task.

Understanding the Impact Copilot for Security Has for Security Professionals *(Ben Edelman, James Bono, Sida Peng, Roberto Rodriguez, and Sandra Ho)*

More details available in Randomized Controlled Trials for Microsoft Copilot for Security (Edelman et al. 2024)

Looking at Copilot in the context of another role, security, researchers extended the lab experiments reported in the first AI and Productivity Report studying Copilot for Security from security novices to security professionals. Participants were recruited through a staffing agency that provides security services to large companies, allowing this new lab study to focus on people who currently use security tools as part of their day-to-day jobs.

Of the 147 security professional participants, three-quarters had 5 or more years of experience as a security analyst. Participants logged into an instance of Microsoft's security service platform, Microsoft Defender, that was created for this experiment. There they performed various tasks, including writing a summary of the incident and answering multiple-choice questions about it. Those with Copilot were 7% more accurate on the multiple-choice questions ($p<.05$). Researchers also asked experts for a list of key facts that should have been included in an incident summary. Study participants with Copilot included 49% more of those key facts in their incident summary reports ($p<.05$). Because it is uninformative to compare speeds across groups when one group is systematically more accurate than another, the researchers looked at quality-adjusted completion times and found that subjects with Copilot were 23% faster overall ($p<.05$).

Compared with the previous study looking at security novices, the security professionals in this study experienced significantly smaller accuracy gains. This is unsurprising given security professionals are more skilled in the tasks and therefore have less room for improvement. Nonetheless, the results show that Copilot allowed professionals to increase their speed without sacrificing accuracy.

Experiment with Licensing Chatbot for Sellers (*Donald Ngwe, Ried Peckham, Ulrike Gruber-Gremlich, and Tyler Smith*)

We next turn to Copilot implications in the sales function. Researchers conducted a lab study to understand how a "licensing chatbot," trained on a corpus of materials around Microsoft's licensing policies, facilitated sellers' ability to answer customer questions. The study asked 64 Microsoft sellers to answer both multiple-choice and open-ended questions in a Qualtrics survey designed to simulate questions that a customer might ask. Sellers were randomly assigned to either have or not have access to the chatbot.

Having the chatbot improved both speed and accuracy. Sellers with the chatbot answered multiple choice questions 3.4 minutes (39%, $p<.05$) faster and accuracy improved by 25 percentage points ($p<.05$). In the open-ended questions, speed, accuracy, completeness, and suitability ratings all improved 34-56% ($p<.05$). These results suggest a positive potential for AI in sales workflows of managing customer sales calls, with potential implications for revenue and customer satisfaction outcomes.

The Effect of Copilot in a Multi-lingual Context *(Benjamin Edelman and Donald Ngwe)*

Another important source of variation is language. Researchers explored Copilot in multilingual contexts, examining how Copilot can facilitate collaboration between colleagues with different native languages.

First, researchers asked 77 native Japanese speakers to review a meeting recorded in English. Half the participants had to watch and listen to the video. The other half could use Copilot Meeting Recap, which gave them an AI meeting summary as well as a chatbot to answer questions about the meeting. Then, researchers asked 83 other native Japanese speakers to review a similar meeting, following the same script, but this time held in Japanese by native Japanese speakers. Again, half of participants had access to Copilot.

For the meeting in English, participants with Copilot answered 16.4% more multiple-choice questions about the meeting correctly, and they were more than twice as likely to get a perfect score. Moreover, in comparing accuracy between the two scenarios, people listening to a meeting in English with Copilot achieved 97.5% accuracy, slightly more accurate than people listening to a meeting in their native Japanese using standard tools (94.8%). This is a statistically significant difference ($p<.05$). The changes are somewhat small in percentage point terms because the baseline accuracy is so high, but Copilot closed 38.5% of the gap to perfect accuracy for those working in their native language ($p<0.10$) and closed 84.6% of the gap for those working in (non-native) English ($p<.05$).

The role of Copilot in communication for non-native speakers has also come up in researchers' interviews with people using Copilot in their day-to-day work. At global companies, Copilot may help people feel more confident that they are communicating effectively. That said, Copilot also raises concerns about potentially increasing the dominance of majority languages: in interviews conducted by other researchers at Microsoft, some people reported changing the language in which meetings were held to one where Copilot was more effective. This effect might shrink or go away as model performance in other languages improves, and improving model performance in non-English languages is a major direction of research at Microsoft and around the world (e.g., Ahuja et al. 2023).

Impact of Generative AI on Metacognition *(Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel)*

More details available in The Metacognitive Demands and Opportunities of Generative AI (Tankelevitch, Kewenig et al. 2024)

Metacognitive demand—the effort needed for monitoring and controlling of one's thoughts and processes—is a part of cognitive load, the total amount of mental effort exerted during tasks. In a review paper (published at the recent ACM SIGCHI 2024 conference) drawing on research in psychology, cognitive science, and the first wave of generative AI lab studies, researchers explored how generative AI (not Copilot specifically) changes the metacognitive demands of a task. For example, prompt engineering, prompt iteration, and output evaluation all require metacognitive effort that may not be needed when doing a task without assistance. This includes work such as developing explicit awareness of task goals, task decomposition, and gaining well-adjusted confidence in one's ability to evaluate the output. Moreover, the availability of generative AI tools creates a more general burden of deciding how to apply these tools to tasks and workflows. To do so "users must have self-awareness of the applicability and potential impact of using GenAI for their workflow; well-adjusted confidence in the ability to complete a task manually versus with GenAI; and metacognitive flexibility in adapting workflows to GenAI."

The researchers suggest that the metacognitive demands of generative AI can be addressed both by improving users' metacognitive abilities and by reducing the metacognitive demands of the tools. Tools could improve metacognitive ability around task decomposition, for example, by responding to a prompt with a list of common subtasks and soliciting (optional) input from the user for each subtask.

The metacognitive demands of AI tools can be further reduced by improving explainability, which can help users calibrate their confidence in outputs and prompting techniques, and self-evaluation or co-auditing. Customizability can either increase or decrease metacognitive demands depending on the context; customizability may make the tool more fitted to the user's task and ability, reducing the metacognitive work of prompting and evaluating, but making customization decisions can also add to cognitive load. As usage spreads and the tools develop, the authors argue that research is needed to understand how tools affect different users' metacognitive processes and what design decisions can lighten the load.

Impact of Copilot on Cognitive Load *(Madeline Kleiner, Max Meijer, Katie Rotella, and Nora Presson)*

One study asked about and tried to measure metacognitive load directly. In this study, 40 Microsoft employees who volunteered for the study created a sales report in Word based on data in an Excel spreadsheet (using a sample report for reference). Half of the participants had access to Copilot and half did not. In addition to measuring time and accuracy, the study asked participants about

how demanding, hard, stressful, and rushed the task was, and the study administered a Stroop test as a measure of participants' cognitive load (Scarpina and Tagini 2017). A Stroop test measures participants response times and error rates in a quick classification task to try to measure the cognitive load they experienced prior to the test.

Participants with Copilot reported the task was less mentally demanding on average (30 out of 100) than the control group (55 out of 100). The improvements for perceived stress and difficulty were similar, with an even larger difference (28 vs. 67 out of 100) for how rushed the task felt. All reported differences have t-test with $p<.05$. It seems that, contrary to the concerns raised in the previous section, in this case the direct help from Copilot counter-balanced or outweighed any increase in metacognitive load. Interestingly, the researchers did not find a difference in the average Stroop score. It is possible that while Copilot made the task feel easier and more enjoyable, it did not affect participants' ability to do the subsequent task. Alternatively, there may have been a small effect on the Stroop score that the study did not have the statistical power to detect, or the Stroop score did not capture the effects observed by the participants in the survey.

4 DISCUSSION

We now look across the studies discussed individually in this report to further explore common themes. Collectively, the studies suggest generative AI is already having positive effects on real-world productivity and begin to highlight the complexity in these effects that emerge from taking a more complete view of work via field research.

As noted above, most previous research in this space relied on lab studies, which by and large provide insight into just a small subset of the tasks people perform as part of their everyday work. The tasks studied in the lab, for example, have tended to require only general knowledge and skills. In contrast, tasks done in the course of work often require highly-specific knowledge and skills. The current generation of generative AI models have been trained primarily on public data, which means they perform best on the types of generic tasks used in these lab studies. Although the gains observed on those tasks in the lab appear to be translating at least to some extent to the real-world tasks that people actually do, it will be important to continue studying AI's impact as models begin to become tailored to perform even better within specific organizational contexts or domains.

More generally, the tasks studied in the lab thus far have tended to be those for which researchers hypothesized generative AI would perform well. This was, in fact, the focus of most of the studies presented in the first AI and Productivity report we published (Cambon et al. 2023). Actual information work, however, often includes a huge variety of tasks and much of the unstructured and informal work in people's jobs is not yet directly supported by the first-generation of generative AI tools. Software developer workflows, for example, involve far more than the hands-on coding supported by GitHub Copilot (Meyer et al. 2017). The ability to shed light on generative AI's productivity dynamics in the natural complexity of entire workflows is a key advantage of field studies of generative AI's productivity impacts, and a major reason we hope to see many more field studies emerging in the literature.

When we look at AI's use in the context of real workflows, we see that context matters a lot. We discussed some initial findings on differences in generative AI usage and effects by an individual's role or function. These findings raise interesting questions in terms of how different roles and functions will find value from generative AI, in terms of efficiencies and also innovation gains. There is an opportunity to further study which individuals and business processes benefit most from AI, and how organizational leaders can enable and encourage AI's productive use. There are also likely many additional sources of heterogeneity, including, for instance, individuals' personalities or the general conditions of the business, e.g. as in Otis et al. (2024).

Assuming generative AI follows the path of most general purpose technologies (Brynjolfsson and McAfee 2014), workflows will, looking forward, be substantially redesigned to better integrate AI. Furthermore, generative AI is still under development and the tools that make use of it are improving rapidly. This means not only that the long-term effects of AI on productivity will differ from those observed in the short-term, but that we are likely to continue to see differences between local task effects and more global productivity effects. Research should try to capture and inform changes in workflows, task design, and business processes in addition to productivity effects for fixed tasks.

One result seen in the above studies and those in our prior work is the common disconnect between the time savings people report from Copilot use and the actual time savings measured. This has been observed not only across studies, where survey measures about time saved tend to be larger than telemetry-based measures, but also within a given study where researchers collect both survey and telemetry measures of time saved on a specific task. There are several potential explanations for these effects, deserving of study. People may enjoy the experience or be excited by the pursuit of the 'answer' with Copilot, which can reduce or speed perceptions of time (Nakamura and Csikszentmihalyi 2002; Gable and Poole 2012). Copilot may also make time appear to go faster as people find it easier to extract and process information (Block et al. 2018; Matthews and Meck 2016), and as people gain experience with Copilot or use more Copilot apps they may perceive increased time savings due to increased ease of use.

An important limitation of the above research and much of the literature on AI and productivity is the near total focus on individual work. The large Early Access Program Telemetry Study described above has some preliminary results on document collaboration, and the researchers are exploring extending their analysis to consider how Copilot affects collaboration networks more broadly, including Outlook and Teams connections. However, given that much of the information work people do is collaborative, it will be

important to further foreground the study of AI's impact on teams and organizations going forward. Additional research is required to understand AI's impact on cross-functional knowledge and cooperation, the social cohesion of teams, and the way information flows across organizations, all of which have implications for growth, productivity and innovation.

5 CONCLUSION

This report provides an overview of the findings from a set of new Microsoft studies that examine the impact of generative AI on information work. It is our second report on the topic, and while the first (Cambon et al. 2023) focused on lab studies, this one focuses on the application of generative AI in real-world workplaces. Across all of the studies discussed, the results suggest that the positive productivity effects that have been observed in a lab setting are beginning to manifest in real-world work. These gains appear to vary contextually (e.g., by role or usage), and these variations indicate there are ways for individuals, organizations, and tool providers to incorporate generative AI in new ways that produce even larger productivity gains for an even wider array of people.

REFERENCES

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K. & Sitaram, S. (2023). MEGA: Multilingual Evaluation of Generative AI. EMNLP 2023.

Anderson, A., Noa Guevara, J., Moussaoui, F., Li, T., Vorvoreanu, M., & Burnett, M. (2024). Measuring User Experience Inclusivity in Human-AI Interaction via Five User Problem-Solving Styles. ACM Transactions on Interactive Intelligent Systems. https://doi.org/10.1145/3663740.

Anderson, L.W. & Krathwohl, D.R. (Eds.) (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Allyn & Bacon. (Pearson Education Group).

Block, R.A., Grondin, S., & Zakay, D. (2018). Prospective and Retrospective Timing Processes: Theories, Methods, and Findings. In Timing and Time Perception: Procedures, Measures, & Applications, pp. 32-51. Brill.

Brynjolfsson, E., Li, D., & Raymond, L.R. (2023). Generative AI at Work. National Bureau of Economic Research. https://www.nber.org/papers/w31161

Brynjolfsson, E. & McAfee, A. (2014). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. 1st edition. W. W. Norton & Company.

Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., Peters, A., & Jernigan, W. (2016). GenderMag: A Method for Evaluating Software's Gender Inclusiveness. In Interacting with Computers, 28(6), 760–787. https://doi.org/10.1093/iwc/iwv046.

Cambon, A., Hecht, B., Edelman, B., Ngwe, D., Jaffe, S., Heger, A., Mihaela Vorvoreanu, M., et al. (2023). Early LLM-Based Tools for Enterprise Information Workers Likely Provide Meaningful Boosts to Productivity. https://www.microsoft.com/en-us/research/publication/early-llm-based-tools-for-enterprise-information-workers-likely-provide-meaningful-boosts-to-productivity/.

Dell'Acqua, F., McFowland III, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Available at SSRN: https://ssrn.com/abstract=4573321.

Doshi, A.R. & Hauser, O. (2023). Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content. Available at SSRN: https://ssrn.com/abstract=4535536.

Edelman, B, Bono, J., Peng, S., Rodriguez, R. & Ho, S. (2024). Randomized Controlled Trials for Microsoft Copilot for Security. Available at SSRN: https://ssrn.com/abstract=4648700.

Gable, P. A., & Poole, B. D. (2012). Time Flies When You're Having Approach-Motivated Fun. Psychological Science, 23(8), 879–886. https://doi.org/10.1177/0956797611435817.

Khemka, M. & Houck, B. (2024). Toward Effective AI Support for Developers: A Survey of Desires and Concerns. Queue 22(3), 53-78. https://doi.org/10.1145/3675416.

Matthews, W. J., & Meck, W. H. (2016). Temporal Cognition: Connecting Subjective Time to Perception, Attention, and Memory. Psychological Bulletin, 142(8), 865.

McMahon, C., Johnson, I., & Hecht, B. (2017). The Substantial Interdependence of Wikipedia and Google – A Case Study on the Relationship Between Peer Production Communities and Information Technologies. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 142-151. https://doi.org/10.1609/icwsm.v11i1.14883.

Meyer, A., Barton, L. E., Murphy, G. C., Zimmermann, T., and Fritz, T. (2017). The Work Life of Developers: Activities, Switches and Perceived Productivity. IEEE Transactions on Software Engineering, 43(12), 1178-1193.

Microsoft. (2024, July 9). AI Data Drop: The 11-by-11 Tipping Point. Retrieved July 9, 2024, from https://www.microsoft.com/en-us/worklab/ai-data-drop-the-11-by-11-tipping-point.

Microsoft & LinkedIn. (2024, May 8). AI at Work Is Here. Now Comes the Hard Part. https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part.

Nakamura, J., & Csikszentmihalyi, M. (2002). The Concept of Flow. Handbook of Positive Psychology, 89-105.

Noy, S., & Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Science, 381(6654), 187-192.

Otis, N., Clarke, R. P., Delecourt, S., Holtz, D., & Koning, R. (2024). The Uneven Impact of Generative AI on Entrepreneurial Performance. Available at SSRN: https://ssrn.com/abstract=4671369.

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv preprint. https://doi.org/10.48550/arXiv.2302.06590.

Rio-Chanona, M., Laurentsyeva, N., & Wachs, J. (2023). Are Large Language Models a Threat to Digital Public Goods? Evidence from Activity on Stack Overflow. arXiv preprint. https://doi.org/10.48550/arXiv.2307.07367.

Scarpina, F. & Tagini S. (2017). The Stroop Color and Word Test. Frontiers in Psychology, 8, 557. https://doi.org/10.3389/fpsyg.2017.00557.

Suri, S., Counts, S., Wang, L., Chen, C., Wan, M., Safavi, T., & Yang, L. (2024). The Use of Generative Search Engines for Knowledge Work and Complex Tasks. arXiv preprint. https://doi.org/10.48550/arXiv.2404.04268.

Tankelevitch*, L., Kewenig*, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The Metacognitive Demands and Opportunities of Generative AI. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-24). https://doi.org/10.1145/3613904.3642902.

Taraborelli, D. (2015). The Sum of All Human Knowledge in the Age of Machines: A New Research Agenda for Wikimedia. ICWSM-15 Workshop on Wikipedia, a Social Pedia: Research Challenges and Opportunities.

Vincent, N. (2022, December 2). The Paradox of Reuse, Language Models Edition. Data Leverage (blog). https://dataleverage.substack.com/p/the-paradox-of-reuse-language-models-edition.

Wiles, E. & Horton, J. (2024). More, but Worse: The Impact of AI Writing Assistance on the Supply and Quality of Job Posts. https://emmawiles.github.io/storage/jobot.pdf

Yeverechyahu, D., & Mayya, R., & Oestreicher-Singer, G. (2024). The Impact of Large Language Models on Open-Source Innovation: Evidence from GitHub Copilot. Available at SSRN: https://ssrn.com/abstract=4684662.