The Heterogeneous Productivity Effects of Generative AI[*]

David Kreitmeir[†]
Paul A. Raschky[‡]

June 9, 2024

Abstract

We analyse the individual productivity effects of Italy's ban on ChatGPT, a generative pretrained transformer chatbot. We compile data on the daily coding output quantity and quality of over 36,000 GitHub users in Italy and other European countries and combine these data with the sudden announcement of the ban in a difference-in-differences framework. Among the affected users in Italy, we find a short-term increase in output quantity and quality for less experienced users and a decrease in productivity on more routine tasks for experienced users.

JEL: D8, J24, O33

Keywords: artificial intelligence, productivity

[*] We would like to thank Yashdeep Dahiya for outstanding research assistance and Samantha Eyler-Driscoll for manuscript editing and proofreading. This paper supersedes a previous version titled "The Unintended Consequences of Censoring Digital Technology – Evidence from Italy's ChatGPT Ban", April 2023.

[†] Department of Economics and SoDa Labs, Monash University (davidkreitmeir.github.io and david.kreitmeir1@monash.edu)

[‡] Department of Economics and SoDa Labs, Monash University (praschky.github.io and paul.raschky@monash.edu)

1 Introduction

The public release of OpenAI's ChatGPT provided near universal[1] access to generative artificial intelligence (AI) tools at no or very low cost. Its subsequent quick adoption[2] broadened the discussion about the impact of generative AI on society and its potential to boost worker productivity by performing relatively complex tasks and producing (seemingly) novel output, all while requiring only minimal technological knowledge on the part of users. However, ChatGPT also has the tendency to produce wrong or faulty outputs (e.g., "hallucinations") that, in the absence of expert knowledge, are difficult to detect and costly to rectify and might ultimately undermine the productivity of some workers (Dell'Acqua et al., 2023).

One of the focal points in the discussion on generative AI's societal impact is its ability to create and generate new content and knowledge. Similarly to prior advances in AI, generative AI can enhance productivity by replacing more routine tasks (Brynjolfsson et al., 2023; Kanazawa et al., 2022; Noy and Zhang, 2023; Peng et al., 2023) or improving users' decision accuracy (Kleinberg et al., 2017; Almog et al., 2024; Cho, 2023). The feature that sets this generation of AI apart from previous ones is that, with its access to the universe of online knowledge, generative AI combines domain-specific information with rules, lending it the ability to create new content and ultimately opening the possibility of extending the production possibility frontier beyond an individual's current level of training or expertise.

However, the accuracy of current generative AI models' performance in some tasks such as text summarization or generation, combined with its clarity and confidence of delivery, might create the *illusion* that it enhances productivity in other domains. Inaccurate, faulty or "hallucinated" output may not be immediately detected and could be used as an input in a knowledge worker's subsequent production flow. For instance, Kabir et al. (2024) analysed ChatGPT's response to 517 programming questions and

[1] Countries where ChatGPT is not accessible include China, Eritrea, Iran, North Korea, Russia and Saudi Arabia, among others.

[2] According to OpenAI, by November 2023, ChatGPT was recording approximately 100 million weekly users. https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/

found that 52% of its answers were incorrect and that the users presented with these answers overlooked these errors 39% of the time. Nevertheless, users still tend to prefer to use ChatGPT because of its comprehensive responses (Kabir et al., 2024) and the confident language of the responses (Li et al., 2023). The biggest online discussion and help forum for developers, stack overflow, has banned the use of LLM generated content on the forum because the rate of "getting correct answers" from these tools, it too low.[3]

For some tasks (e.g., content writing), the process of detecting faulty output or rectifying generative AI–driven errors might be quick, while for others (e.g., software development), the same process can be tedious and time consuming.[4] There are also wide differences in the accuracy of generative AI output, driven not only by the complexity of the underlying task but also the size and quality of the underlying training data. The tools can leverage a very large online text corpus to predict the next word in tasks such as creative writing and chatting, but the training data for software development and code creation are limited to a relatively small number of online forums (e.g., Stack Overflow), where the ground truth can be noisy.[5]

In cases where the underlying task is more complex and the output requires accuracy to be ultimately useful (e.g. software development), relying on generative AI might prolong task completion and decrease workers' output quality (Dell'Acqua et al., 2023). Such problems might be more acute among less experienced workers, who may have less domain knowledge and require more time to detect and correct errors. Less experienced workers might also be more prone to continue using the tool because the alternatives

[3] https://meta.stackoverflow.com/questions/421831/policy-generative-ai-e-g-chatgpt-is-banned

[4] While the motivation for the relatively early public release of ChatGPT and other large-language models (LLMs) was to improve their performance with the human-generated data collected from user interactions with the tools, their output quality remains noisy, and expert knowledge is often required to accurately judge this quality. Moreover, Chen et al. (2023) show that ChatGPT's performance on a number of tasks, including generating code, actually declined from version 3.5 to version 4.0, calling into question whether its performance and accuracy will continuously improve over time. In addition, del Rio-Chanona et al. (2023) show that the widespread use of ChatGPT has led to a decline in usage of online help forums such as Stack Overflow, which in return will decrease the human-generated ground truth data that can be used to improve AI models.

[5] For example, for less routine, more complex and more niche questions, the answers provided on Stack Overflow are not necessarily correct or are just initial solution suggestions instead of working solutions. While these suggestions might have received upvotes, signaling to the LLM their "usefulness" for its training, in reality, the content of the answers and ChatGPT subsequent output might be only an untested and ultimately not functioning code routines.

(e.g., acquiring the knowledge and skills themselves) appear to be even costlier.

In this paper, we use observational data to analyse the heterogeneous effects of ChatGPT on the output quantity and quality of experienced and less experienced software developers. In particular, we exploit Italy's sudden announcement of a ChatGPT ban as a natural experiment to examine the ban's short-run effects on GitHub users' productivity. We find that the ban had no systematic effect on the overall output of more experienced developers and only some small negative effects on their output for more routine tasks (resolving issues and debugging). However, among less experienced users, the short-term lack of access to ChatGPT increased both the amount of output and its quality. For this group of users, the likelihood that we observe any output-related activity on GitHub is approximately 10% higher for the two business days following the ban. This effect size shrinks for the subsequent days. In the same vein, we find some tentative evidence that Internet users in Italy adapted fairly quickly to the legislation by increasing their use of virtual private networks (VPNs) and encrypted routing to circumvent the ban.

Finding high-frequency and consistently measured, observational data on the output of knowledge workers that is comparable across countries is challenging. We follow a the approach from a number of recent papers (e.g., McDermott and Hansen, 2021; Holub and Thies, 2023; Shen, 2023) that have already exploited the temporal granularity of GitHub activity data and used this data as a proxy for software developers' productivity. When using this data to proxy productivity, we rely on a number of assumptions which we test in a subsequent robustness section. We show that the results are not driven by changes in working hours or the level of complexity of the task. Considering that the ban was implemented close to a major holiday in the treatment and control countries, we also show that the changes in behaviour are not driven by seasonal factors. Our results are also robust to the use of a set of alternative outcome variables and also analysing the effect at the user–repository–day level.

Our results present some first, nonexperimental empirical evidence on the effects of restricting access to generative AI on workers' performance in more complex tasks. Importantly, we show that the effects of generative AI are heterogeneous by worker's experience.

Our study complements the existing, largely experimental, literature on the effects of generative AI on worker productivity in less complex tasks (e.g., content writing, customer support), where generative AI output is less error-prone, by examining a setting with more complex tasks, where AI-generated output can be less accurate or more faulty. Existing work by Brynjolfsson et al. (2023) and Noy and Zhang (2023) has found mainly positive productivity effects of generative AI and stronger effects for less experienced workers in the contexts of customer support and content writing tasks. In contrast, our results suggest that, for more complex tasks (e.g., code development),[6] generative AI does not necessarily boost the productivity of less experienced workers and can even decrease their output quantity and quality. Our results, therefore, confirm the findings of Dell'Acqua et al. (2023) in a controlled experiment environment that, for tasks beyond ChatGPT's current capabilities, using ChatGPT increases the time a worker spends on a task. We complement their results by highlighting the differential effects by knowledge worker's level of experience. Our finding of heterogeneous effects for more complex tasks also empirically complements the larger discussion in economics on technological change and inequality in the labour market (e.g., Acemoglu, 2002; Autor et al., 2003; Goldfarb and Tucker, 2019; Acemoglu and Restrepo, 2020), in particular the productivity and labour market effects of AI (e.g., Brynjolfsson et al., 2017; Agrawal et al., 2019; Acemoglu, 2021; Eloundou et al., 2023).

The paper is organised as follows: Section 2 provides some background on ChatGPT and the Italian ban on the technology in 2023. Section 3 describes the data. Section 4 presents empirical results on the ban's effect on worker productivity, and Section 5 concludes.

[6] Related work by Peng et al. (2023) and Chatterjee et al. (2024) has found positive effects of GitHub Copilot on both the productivity and job satisfaction of software developers. While GitHub Copilot is also an AI-based tool developed by GitHub and OpenAI, it is specifically a code completion tool. In contrast, ChatGPT is a chat-based (rather than auto-complete) tool and can be used to produce entirely new code segments/programs based on a human language prompt; in such cases, the accuracy of generative AI is highly variable (Kabir et al., 2024).

2 ChatGPT and the Italian Ban

ChatGPT, an LLM created by US startup OpenAI, has been used by millions of people since it launched in November 2022. Trained on a vast corpus of text data from the Internet as it was in 2021, this large-scale AI language model uses a transformer-based neural network to process natural language. During the training process, the model learned to identify patterns and relationships between words, phrases, and sentences, enabling it to generate text.

ChatGPT is accessible via a public website (chatgpt.openai.com) or an application programming interface (API), and almost anyone[7] can sign up for a free account. The interface is designed like a chat environment where the user writes "prompts" and ChatGPT answers. Interactions can range from casual chats and search-like queries to more complex exchanges such as creative writing of a text or creation of recipes based on prompts. ChatGPT can also write code in multiple programming languages on the basis of a simple prompt.

On April 1, 2023, the Italian data protection authority (Garante per la protezione dei dati personali) blocked use of the ChatGPT chatbot, citing privacy concerns, and announced an investigation into OpenAI's compliance with the European Union's General Data Protection Regulation (GDPR). In particular, the authority stated that there was no legal basis for the mass collection and storage of personal data to train the algorithms underlying the platform's operation.[8] There are a number of reasons to expect that the ban did not have an effect on software developers. During the ban access to other Natural Language Processing (NLP) powered tools that are powered by Open AI, such as GitHub CoPilot, was not affected. In contrast to ChatGPT which was trained on the universe of available online text corpus, GitHub Copilot was specifically trained only on code-repositories and build to produce context-aware codes. It provides more accurate results and is also widely used by code-developers. It is important to note, that our study only estimates the effects of a ban of ChatGPT on developer's output and not a ban on

[7] Before Italy, countries including China, Russia and North Korea had already banned ChatGPT.

[8] Shiona McCallum, "ChatGPT banned in Italy over privacy concerns", BBC 01/04/2023, https://www.bbc.com/news/technology-65139406

all LLM based coding support tools. Alternative generative AI tools such as Claude and Bard which were launched a few weeks prior to the ban were still accessible. There were also multiple ways for programmers to immediately circumvent the ban (i.e. VPN) and guides on how to do that where shared online.[9]

However, anecdotal evidence shows that shortly after the inception of the ban, Italian developers went online and complained about the disruption caused by restricting access to ChatGPT, "[...] a tool that has become an essential part of [their] daily routine" as software developers.[10] Concerns were raised that restricting access to ChatGPT during a time when the field of software development is increasingly fast-paced, poses a severe threat to the competitiveness Italian developers and businesses.[11] Whether the ban was effective or not is ultimately an empirical question, which this paper aims to shed light on.

The ban was lifted in late April after OpenAI responded to the data protection authority's privacy concerns.[12]

3 Data

GitHub Data GitHub is the world's largest online code hosting platform, used for storage of and joint work on coding projects (so-called repositories).[13] All modifications to a GitHub repository are automatically timestamped and stored, and GitHub permits tracking of any iterations of specific files and lines of code. Every action taken by a team member is automatically recorded, with details about the kind and substance of the modification, the files and code lines affected, and the date the changes were performed. Anyone with access to a repository can examine and download the history of iterations

[9] https://www.programmareinpython.it/blog/chatgpt-bloccato-in-italia-che-fare/

[10] Semero, Anto "ChatGPT Banned in Italy: Mamma Mia! What's Going On?" https://medium.com/@antonellosemeraro/chatgpt-banned-in-italy-mamma-mia-whats-going-on-97c44284e331, April 2 2023.

[11] https://www.programmareinpython.it/blog/chatgpt-bloccato-in-italia-che-fare/

[12] Shiona McCallum, "ChatGPT accessible again in Italy", BBC 28/04/2023, https://www.bbc.com/news/technology-65431914

[13] The programming languages most commonly represented in GitHub repositories are `Python` (17.38%), `Java` (11.77%), `Go` (10%), `JavaScript` (9.95%), and `C++` (9.66%). In comparison, `R`-related repositories account for only 0.074% of all pull requests on GitHub. https://madnight.github.io/githut

and actions, and given GitHub's history of developing open-source software, a significant portion of its repositories are not access restricted, meaning that the project activity information is available to everyone. Thus, public GitHub repositories provide a direct, real-time measure of labour activity for millions of software and code developers worldwide (McDermott and Hansen, 2021).[14]

We access individual-level, real-time activity data from the GitHub Archive for GitHub users in Italy (treatment) and Austria, France, and Spain (control) in the week prior to and that immediately after the ChatGPT ban in Italy (March 27–April 9, 2023).[15] To account for the Easter break starting with Good Friday on 7 April—a public holiday in all four sample countries— we restrict the post-treatment period to Monday 3 April to Thursday 6 April and the corresponding pre-period to Monday 27 March to Thursday 30 March.[16] GitHub Archive is hosted on Google's BigQuery warehouse system and contains all public event data of GitHub user, which is updated daily and can be accessed with a query on Google's cloud infrastructure. GitHub user information such as the year of GitHub user account creation was downloaded with the GitHub GraphQL API.[17] The two datasets are merged via the unique GitHub user login.

We use the individual-level action data to construct two sets of baseline outcome variables: The first group captures *output quantity and quality* and includes aggregate *Output* limited to "productive" actions, aggregate *Output* as defined by Shen (2023), aggregate *Output* as defined by Holub and Thies (2023), and the *Pull request (PR) merge ratio* quantifying how many of a user's suggested code edits were accepted by the repository (project) owners. The second group gauges *task choice and complexity* and

[14] GitHub data have been used in empirical research on software developers' productivity during the onset of COVID-19 (Forsgren, 2021), the impact of COVID-19 on daily and weekly patterns of individual labour allocation (McDermott and Hansen, 2021), the effects of working from home on individual productivity (Shen, 2023), the effect of air pollution on individual output (Holub and Thies, 2023), and the relationship between social links and the likelihood of joining professional software development teams (Casalnuovo et al., 2015).

[15] We chose these three countries because all of them are part of the European Union and share a common land or sea border. In settings like ours, it is difficult to find objective criteria that help guiding the choice of comparable units. To provide further confidence that our results are not based on the choice of control countries, we conduct a leave-one-out analysis in Figure C.6 and show that the results are not driven by the composition of the control group.

[16] For more details on the study's time line and a graphical illustration please see A in the appendix.

[17] For more information on how we retrieve GitHub user location information, please see B.1 in the appendix.

comprises *PR opened*, *Avg. lines added per opened PR*, *Avg. lines added per merged PR*, *Easy issue closed*, and *Interactive activity.* A detailed description of the construction and definition of all outcome variables is provided in appendix Table B.1.

On a daily basis, GitHub actions are relatively rare events at the user level. Hence, we transform each count variable into a binary indicator that equals 1 if one of the actions in a category is recorded for the user on a given day, and 0 otherwise.[18] Descriptive statistics at the user–day and user level are presented in appendix Table D.1. To use GitHub actions as a measure for developers' productivity, we need to ensure that the ban did not affect developers' working hours. While it is not possible to have data on exact working hours for developers, we are still able to check if time spent on GitHub activity during the day has changed between the pre and post period. Using the timestamp data for each individual GitHub action, we are able to show that the distribution of GitHub activities across hours of the day did not change over the period (see Figure C.3).

Package Repositories We compile a list of packages hosted on GitHub for ten analytical programming languages: `C`, `C++`, `Go`, `Java`, `JavaScript`, `Julia`, `Perl`, `Python`, `R`, and `Rust`. We rely in the first instance on the community–curated "Awesome Lists" to locate GitHub repositories for "popular" packages in each language. In a second step, we scrape the information on all packages hosted on the official software repositories for `Python` (`pypi`), `R` (`CRAN`) and `Julia` (`JuliaRegistries`) to retrieve information on each package's GitHub repository. We make use of the standardized GitHub URL structure to identify the *owner* and *name* of a package repository.[19] To identify the GitHub user accounts other than the owner that contribute to a package repository, we use information on each individual GitHub user's activity from January 2011 until March 2023. We restrict the list of *GitHub event types* to "productive" events to select primarily accounts that made at least one substantial contribution to a package repository.[20] Moreover,

[18] The distribution of day–user-level counts of the main event variables during the sample period is presented in Figure C.1 in the appendix.

[19] The stylized URL for a package repository is `https://GitHub.com/[account name hosting the repository]/[repository name]` (for instance, `https://github.com/numpy/numpy`).

[20] Our set of "productive" event types comprises `PullRequestEvent`, `PullRequestReviewEvent`, `PullRequestReviewCommentEvent`, `PushEvent`, and `ReleaseEvent`.

we winsorise the sample at the 1st and 99th percentiles to safeguard against outliers.[21] Our final list of package contributors and owners comprises $483,855$ unique GitHub user accounts, of which $5,916$ are part of our baseline sample.

4 Effect of the ChatGPT Ban on GitHub Output

To analyse the effect of the Italian ChatGPT ban on GitHub users' output, we estimate variants of the following difference-in-difference (DID) specification:

$$Y_{it} = \beta D_{it} + \alpha_i + \lambda_t + \gamma_{dow} + \sigma_d \times t + \epsilon_{it}, \tag{1}$$

where Y_{it} denotes the outcome variable, i.e. one of the user-specific output and task variables, and D_{it} is the treatment indicator variable equaling 1 for Italian users in the first four work days post the ChatGPT ban. The parameters of interest is β. Time-invariant differences between users, including ability and experience, are captured by user fixed effects α_i, while day (date) fixed effects λ_t account for daily fluctuations in coding output across users. Additionally, we account for differences in working behaviour across week-days via day-of-the week γ_{dow} and include control treatment and control group specific time trends to safeguard against differential time-trends in coding activity between Italian Github users and their European peers.[22]

To check for potential pre-trends and investigate how the estimated effect evolves over time, rather than averaging over the whole window as in the generalized DID specification, we also estimate the following event-study specification:

$$Y_{it} = \sum_{\tau=-4}^{-2} \beta_\tau D_{it}^\tau + \sum_{\tau=0}^{3} \beta_\tau D_{it}^\tau + \alpha_i + \lambda_t + \gamma_{dow} + \sigma_d \times t + \epsilon_{it}, \tag{2}$$

where D is a dummy variable equalling one for observations in the treatment group at event–day τ and zero otherwise; with $\tau = -1$ serving as the (excluded) reference period.

[21] Note that we exclude `bot` accounts from the list of contributors prior to winsorising.

[22] We estimate alternative specifications of our baseline regression model. Including country-specific linear time-trends or excluding time-trends altogether leaves our baseline estimates quantitatively and qualitatively stable (Table C.2 and Table C.3 in the appendix).

In both specifications, we cluster standard errors at the user level.

4.1 Baseline Results

Table 1 presents our baseline DID model estimates for the coefficient of interest $\widehat{\beta}$. The upper panel reports the average treatment effect of the ChatGPT ban on output quantity and quality for four different samples (overall, less experienced, experienced and package contributors), while Panel B presents baseline estimates for task choice and complexity. We find that the ban has *overall* no significant effect on the output quantity and quality of Italian users or their choice of tasks with the notable exception of a negative and significant impact on their ability to close issues. These aggregate results, however, disguise important heterogeneity across users.

In Panel B, we provide suggestive evidence that the estimated effect on output quantity and quality of *less experienced* users is not result of a change in tasks. In particular, we do not see that the number of pull requests send for review (column 1) changes nor their complexity (column 2). Further, restricting the focus to lines added on pull requests that were actually merged (column 3), reveals a significant positive effect. While this increase could reflect less efficient code by Italian users post-ban, our estimates provide no support for the notion that the increase in the PR merge ratio is the result of decrease in merged PR task complexity. Similarly, we do not observe a switch to easier issues (column 4) nor to collaborative engagement instead of coding activity (column 5).[23] For *experienced* users, we find null effects of the ban on task allocation and complexity, while the ban appears to have some effects on *package contributors*. In particular, the decline in the probability of opening a pull request and resolving easy issues might indicate that package contributors tackle less but harder tasks after the ban.

Figure 1 presents the estimated event-study coefficients β_τ from specification 2 for *less experienced* users. Importantly, a joint F-test of whether all coefficients prior to the

[23] In unreported results, we test if the estimated effect on *easy* issues is the result of restricting the attention to the resolution of those issues. If we consider any form of engagement with easy issues—i.e. commenting, opening or closing—we find that the probability of engaging with easy issues declines for less experienced users (significant at the 10% level), while no change is observed for more sophisticated users.

ban are jointly zero cannot be rejected at conventional levels for any outcome, alleviating concerns about preexisting trends. Further, the event-study results reveal that the treatment effect for less experienced users peaked two days after the ban. Figure C.4 and C.5 in the appendix present the corresponding event-study estimates for *experienced* users and *package contributors*, respectively. Results for this subset of users reveal that point estimates are rather volatile over time, particularly for *package contributors*, and reveal no clear pattern except for *issue closed.* Overall, our event-study results provide support for our main DID findings that the ban actually increased the productivity of *less experienced* coders, while more sophisticated users were largely unaffected.

4.2 Robustness Checks

We conduct a number of robustness checks and present the results in the appendix.

Alternative outcomes: First, we consider a number of additional outcomes in Panel A of Table C.4. We show that our results are largely robust to alternative measures for output quantity (*Any Event* and *Commit*), output quality (*PR merged* and *PR merge ratio (Holub and Thies, 2023)*) as well as task complexity (*Avg. files edited per merged PR*). In Panel B, we additionally show that our baseline findings are qualitatively stable at the intensive margin when we use continuous outcome variables instead of binary indicators.

Placebo tests: Second, the overall changes in output quantity and quality could be driven by unobserved factors occurring at the same time as the introduction of the ban or general differences in Italian Github user activity relative to their European peers in the work week leading up to the Easter break. To address these concerns, we undertake two placebo exercises. In the first placebo test, we assume that the ban was implemented on the weekend prior to our *actual* pre-treatment period from Monday 27 to Thursday 30 March 2023. The estimated *placebo* treatment effects are presented in Table C.5 and are relatively small in magnitude and vastly insignificant. Importantly, we don't find the heterogeneous effects across users uncovered in our baseline analysis. Our second *placebo* treatment period spans the work week prior to Good Friday on 15 April 2022. Reassuringly, our placebo estimates (Table C.6) are again exceedingly insignificant and do not follow a clear and comparable pattern to our baseline results.

Repository-level Analysis: For a subset of users working on multiple repositories, we construct a new panel dataset at the user–repository–day level (Section D in the appendix).[24] This allows us to include repository (project) and user fixed effects. Exploiting only within user-repository variation projects we estimate the effect of the ban on users while holding the complexity of a project constant. The results in Table D.2 largely confirm the patterns from our baseline, user-level, analysis. For less experienced users, we still observe a positive and significant effect on output quantity and quality, while the likelihood of closing an issue for is still negatively affected by the ban. Interestingly, we now also find some suggestive evidence for the negative impact of the ban on experienced users' output quantity relative to that of their European peers working on the same repository. Moreover, we find additional support for the negative effect of the ban on experienced users when conditioning on programming language and restrict our attention to highly complex official package repositories (Table D.3).[25]

Additional Robustness Checks We conduct a number of additional sensitivity checks. First, we conduct a "Leave-One-Out" analysis to further alleviate concerns that our baseline findings are dependent on our selection of control group countries. In particular, we show in Figure C.6 that our results are robust to excluding users from each of the control group countries. Finally, we address concerns that our results could suffer from an over-rejection of the null hypotheses as a result of reuse of the identifying exogenous variation for multiple outcomes (and subsamples). Table C.8 reports p-values (in brackets) that are corrected for multiple hypothesis testing following the procedure described in Romano

[24] Note that the set of public repositories is also restricted to those with at least one user from both the control and treatment group.

[25] Because of the limited number of observations for package repositories of other programming languages, our analysis is, unfortunately, constrained to official `python` packages from `PyPi`. In addition, we exclusively present results for output quantity and quality in D.3 due to the limited variation in the data for most outcomes concerning task choice and complexity.

and Wolf (2005a,b, 2016).[26] Importantly, output quantity as defined in Shen (2023) and, in particular, output quality retain their significance even after correcting for multiple hypothesis testing.

4.3 Discussion

Data from public GitHub repositories include code and software projects from a variety of organisations and individuals. Some of these are open-source development projects (e.g., APIs) from private-sector companies, some are general open-source projects developed by a community of volunteers (and therefore are closer in character to public goods), and others are owned by research organisations or individual developers. Given data limitations, it is not possible to distinguish the type of project.

It is possible that some Italian users immediately used tools (e.g., VPNs) to circumvent the ban. Using data on Google searches for VPN services and usage data for TOR[27] (see appendix E), we show a sudden jump in circumvention activity among Italian Internet users in the days after the ban. Despite the easy access to circumvention technology, many corporations and organisations actually prohibit the use of VPN and TOR tools on their devices and networks, implying that their use may be limited to mainly private devices and home networks. More importantly, we still find systematic effects on output despite this circumvention activity, and one can interpret our results as a lower bound. Another concern is that our finding of heterogeneity between less experienced and experienced users could be driven by the latter's greater skill in circumventing the ban. However, the systematic effects of the ban on tasks related to closing issues and the negative output effect detected in repository-level analysis suggests that this is not the case.

There are a number of follow-up questions that we are unable to empirically analyse because of limitations in the data. First, while generative AI might disrupt the production flow of less experienced workers by providing faulty results, another possibility is that

[26] In recent work, Heath et al. (2022) show that the employed *Romano–Wolf* correction procedure performs well in a multitude of settings and across different dimensions.

[27] The TOR (The Onion Router) network is an open-source overlay network of thousands of network relays that conceals a user's IP address. Unfortunately, we cannot access actual VPN usage data at daily level.

ChatGPT is simply a distraction. While it is not clear how this would explain the effect heterogeneity, more detailed data on the actual use of ChatGPT could help inform the design of workplace policies around generative AI.[28] Second, more detailed data would also shed light on the question of why, after the initial increase in output and quality, we observe a decrease in the effect size for less experienced users in subsequent days. One explanation, in line with the conclusions of Kabir et al. (2024) and Li et al. (2023), could be that less experienced users still prefer to use ChatGPT as a support tool because it generates accessible and easy-to-use responses and because the costs of pursuing alternative solutions (e.g., acquiring the necessary coding skills) is relatively high. Finally, our study provides evidence on the productivity effects of (the ban on) generative AI in only the very short run because the ban was short-lived and circumventing it was relatively easy.

5 Conclusion

We present novel evidence of the short-term effects of generative AI (ChatGPT) on the productivity of knowledge workers using high-frequency, observational data from over 36,000 software developers in Italy and other European countries. We use the sudden ban on ChatGPT in Italy as a natural experiment and show that the access restriction distorted output quantity and quality. Our results not only present some first empirical evidence of the widespread adoption of ChatGPT in software and code development but also show that the productivity effects of ChatGPT (and restrictions on it) differ by experience level. Our findings have the following policy implications: For some, more complex tasks, generative AI can produce faulty and erroneous output that is difficult to detect, in particular for less experienced individuals. This calls for a more targeted use of the tool in both education and work. AI-based tools that harness the power of LLMs in a more controlled form, that generate a clearly defined output and that are not based on simple text prompts (e.g., GitHub Copilot) offer guard rails to ensure more domain-specific use. Our

[28] For example, existing generative AI tools such as GitHub Copilot are in general productivity enhancing because they are designed for specific tasks; Copilot, for example, only completes code and is not an open-ended chatbot.

findings also indicate that even well-intended government-mandated blocking of digital technology (to protect privacy) can lead to short-term output disruptions and costs for society. Sudden bans can be easily circumvented with VPN tools, but these adjustment activities simultaneously distort production processes and negatively impact productivity in professions that rely on the banned technology. Thus, our research also implies that policymakers should consider the potential economic cost of digital technology bans before imposing them.

References

Acemoglu, Daron, "Technical change, inequality, and the labor market," *Journal of Economic Literature*, 2002, *40* (1), 7–72.

_ , "Harms of AI," Working Paper 29247, National Bureau of Economic Research September 2021.

_ **and Pascual Restrepo**, "Robots and jobs: Evidence from US labor markets," *Journal of Political Economy*, 2020, *128* (6), 2188–2244.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb, "Artificial Intelligence: The ambiguous labor market impact of automating prediction," *Journal of Economic Perspectives*, May 2019, *33* (2), 31–50.

Almog, David, Romain Gauriot, Lionel Page, and Daniel Martin, "AI oversight and human mistakes: Evidence from Centre Court," arXiv 2401.16754 2024.

Autor, David H., Frank Levy, and Richard J. Murnane, "The skill content of recent rechnological change: An empirical exploration," *Quarterly Journal of Economics*, 11 2003, *118* (4), 1279–1333.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics," Working Paper 24001, National Bureau of Economic Research November 2017.

_ **, Danielle Li, and Lindsey R Raymond**, "Generative AI at work," Working Paper 31161, National Bureau of Economic Research April 2023.

Casalnuovo, Casey, Bogdan Vasilescu, Premkumar Devanbu, and Vladimir Filkov, "Developer onboarding in GitHub: The role of prior social links and language experience," in "Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering" ESEC/FSE 2015 Association for Computing Machinery New York, NY, USA 2015, p. 817–828.

Chatterjee, Sayan, Ching Louis Liu, Gareth Rowland, and Tim Hogarth, "The impact of AI tools on engineering at ANZ Bank: An empirical study on GitHub Copilot within corporate environment," arXiv 2402.05636 2024.

Chen, Lingjiao, Matei Zaharia, and James Zou, "How is ChatGPT's behavior changing over time?," arXiv 2307.09009 2023.

Cho, Sungwoo, "The effect of robot assistance on skills," mimeo, UCLA Economics 2023.

del Rio-Chanona, Maria, Nadzeya Laurentsyeva, and Johannes Wachs, "Are Large Language Models a threat to digital public goods? Evidence from activity on Stack Overflow," arXiv 2307.07367 2023.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "The diffusion of disruptive technologies," Working Paper 24-013, Harvard University 2023.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: An early look at the labor market impact potential of Large Language Models," 2303.10130, arXiv 2023.

Forsgren, Nicole, "An analysis of developer productivity, work cadence, and collaboration in the early days of COVID-19," Technical Report, Octoverse Spotlight 2021.

Goldfarb, Avi and Catherine Tucker, "Digital Economics," *Journal of Economic Literature*, March 2019, *57* (1), 3–43.

Heath, Davidson, Matthew C. Ringgenberg, Mehrdad Samadi, and Ingrid M. Werner, "Reusing Natural Experiments," *Journal of Finance*, 2022, *Forthcoming.*

Holub, Felix and Beate Thies, "Air quality, high-skilled worker productivity and adaptation: Evidence from GitHub," CRC TR 224 Discussion Paper Series crctr224_2023_402, University of Bonn and University of Mannheim, Germany March 2023.

Kabir, Samia, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang, "Is Stack Overflow obsolete? An empirical Sstudy of the characteristics of ChatGPT answers to Stack Overflow questions," arXiv 2308.02312 2024.

Kanazawa, Kyogo, Daiji Kawaguchi, Hitoshi Shigeoka, and Yasutora Watanabe, "AI, skill, and productivity: The case of taxi drivers," Working Paper 30612, National Bureau of Economic Research 2022.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 08 2017, *133* (1), 237–293.

Li, Jiachen, Elizabeth Mynatt, Varun Mishra, and Jonathan Bell, ""Always nice and confident, sometimes wrong": Developer's experiences engaging Generative AI Chatbots versus human-powered Q&A platforms," arXiv 2309.13684 2023.

McDermott, Grant R and Benjamin Hansen, "Labor reallocation and remote work during COVID-19: Real-time evidence from GitHub," Working Paper 29598, National Bureau of Economic Research December 2021.

Noy, Shakked and Whitney Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, 2023, *381* (6654), 187–192.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, "The impact of AI on developer productivity: Evidence from GitHub Copilot," arXiv 2302.06590 2023.

Romano, Joseph P and Michael Wolf, "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 2005, *100* (469), 94–108.

_ **and** _ , "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 2005, *73* (4), 1237–1282.

_ **and** _ , "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing," *Statistics & Probability Letters*, 2016, *113*, 38–40.

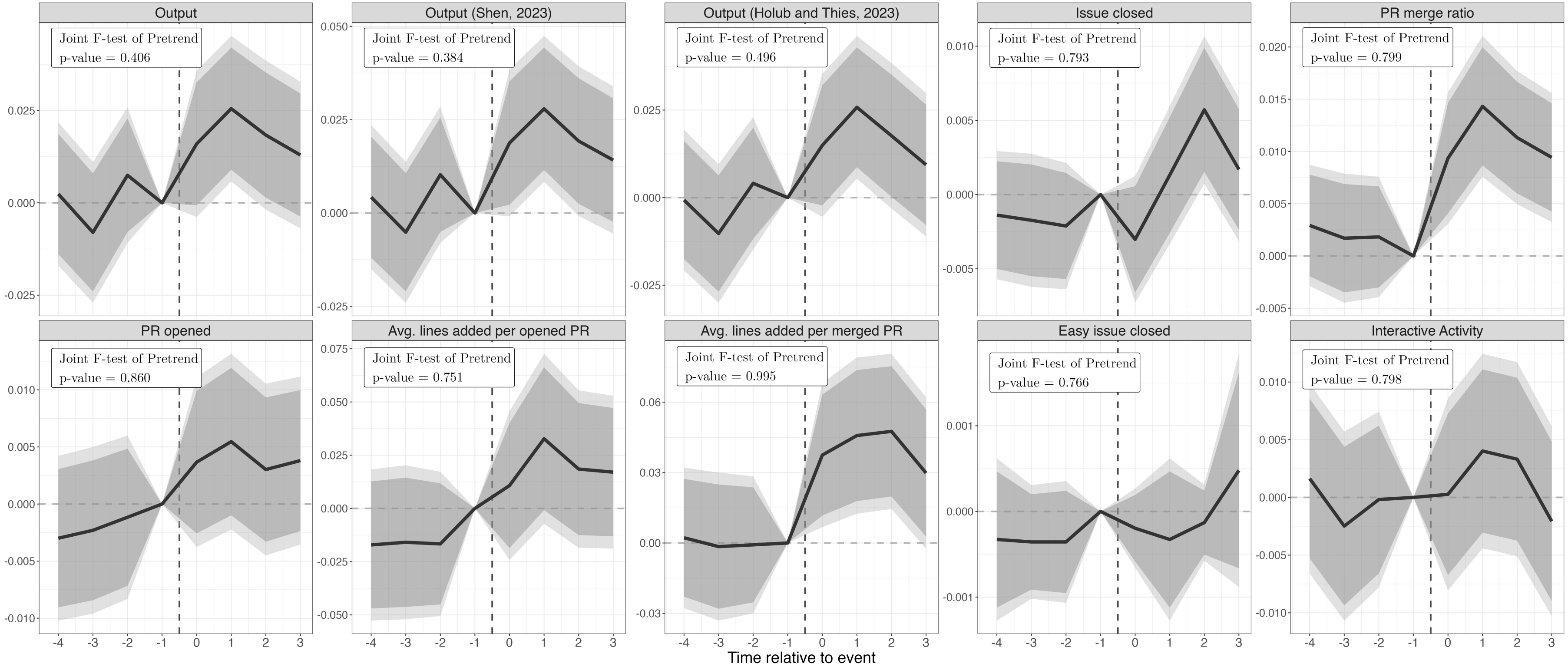
Shen, Lucas, "Does working from home work? A natural experiment from lockdowns," *European Economic Review*, 2023, *151*, 104323.

Table 1: Effect of ChatGPT Ban on GitHub Output

		(1)	(2)	(3)	(4)	(5)
A: Output Quantity and Quality						
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023)	Issue closed	PR merge ratio
Overall	Treated × Post	0.0062	0.0102	0.0090	-0.0045**	0.0045
(N = 290,864)		(0.0076)	(0.0075)	(0.0080)	(0.0022)	(0.0030)
	Dep. var. mean	0.2314	0.2283	0.2624	0.0142	0.0359
Less experienced	Treated × Post	0.0193*	0.0216**	0.0203*	-0.0017	0.0118***
(N = 149,680)		(0.0107)	(0.0107)	(0.0111)	(0.0023)	(0.0035)
	Dep. var. mean	0.2339	0.2316	0.2519	0.0089	0.0258
Experienced	Treated × Post	-0.0079	-0.0020	-0.0030	-0.0077*	-0.0034
(N = 141,184)		(0.0107)	(0.0106)	(0.0116)	(0.0039)	(0.0052)
	Dep. var. mean	0.2287	0.2249	0.2734	0.0198	0.0467
Pkg. contributor	Treated × Post	0.0181	0.0164	0.0270	-0.0151*	0.0113
(N = 47,328)		(0.0205)	(0.0200)	(0.0215)	(0.0084)	(0.0104)
	Dep. var. mean	0.2688	0.2644	0.3312	0.0267	0.0615
B: Task Choice and Complexity						
		PR opened	Avg. lines added per opened PR	Avg. lines added per merged PR	Easy issue closed	Interactive Activity
Overall	Treated × Post	0.0000	-0.0039	0.0220	-0.0004	0.0006
(N = 290,864)		(0.0034)	(0.0148)	(0.0135)	(0.0004)	(0.0043)
	Dep. var. mean	0.0402	0.1581	0.1439	0.0007	0.0637
Less experienced	Treated × Post	0.0040	0.0211	0.0456***	-0.0004	0.0037
(N = 149,680)		(0.0040)	(0.0186)	(0.0164)	(0.0004)	(0.0045)
	Dep. var. mean	0.0315	0.1358	0.1148	0.0005	0.0340
Experienced	Treated × Post	-0.0046	-0.0324	-0.0031	-0.0003	-0.0028
(N = 141,184)		(0.0056)	(0.0237)	(0.0221)	(0.0008)	(0.0077)
	Dep. var. mean	0.0494	0.1817	0.1749	0.0009	0.0953
Pkg. contributor	Treated × Post	-0.0221*	-0.0933*	0.0226	-0.0021*	0.0127
(N = 47,328)		(0.0117)	(0.0477)	(0.0422)	(0.0011)	(0.0158)
	Dep. var. mean	0.0661	0.2356	0.2210	0.0014	0.1393

Notes: All specifications include user–fixed effects, day-of-the-week–fixed effects, and a linear time trend for the control and treatment group. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Robust standard errors in parentheses are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1: Less Experienced Users – Event-Study Estimates



Notes: Event-study estimates across outcomes for "less experienced" GitHub user accounts (created after or in 2017). The sample period spans March 27–30 (*Pre*) and April 3–6 (*Post*). All specifications include user, time, and day-of-the-week fixed effects. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). 95% (90%) confidence intervals for robust standard errors clustered at the user level are depicted in light (dark) grey.

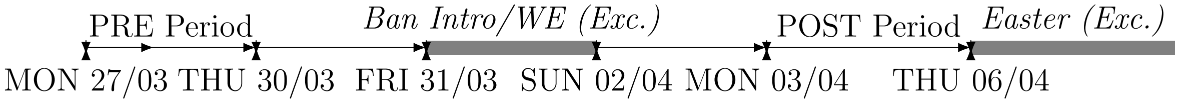
Supplementary Online Appendix

Table of Contents

A Timeline

Figure A.1 depicts the timing of the events. The ban was introduced on Friday March 31st 2023. In our baseline analysis, we consider the 4 business days prior to the ban, (MON – THU, March 27th – 30th) as the pre-period. To make the analysis, comparable, we ignore the two days following the introduction of the ban which were weekend days (SAT and SUN, April 1st – 2nd, in gray are *excluded*). The Post period is defined as the first 4 business days, after the introduction of the ban (MON – THU, April 3rd – 6th 2023). We further exclude Friday, 7th 2023, because this was Good Friday, a national holiday in all European countries under consideration in our sample.

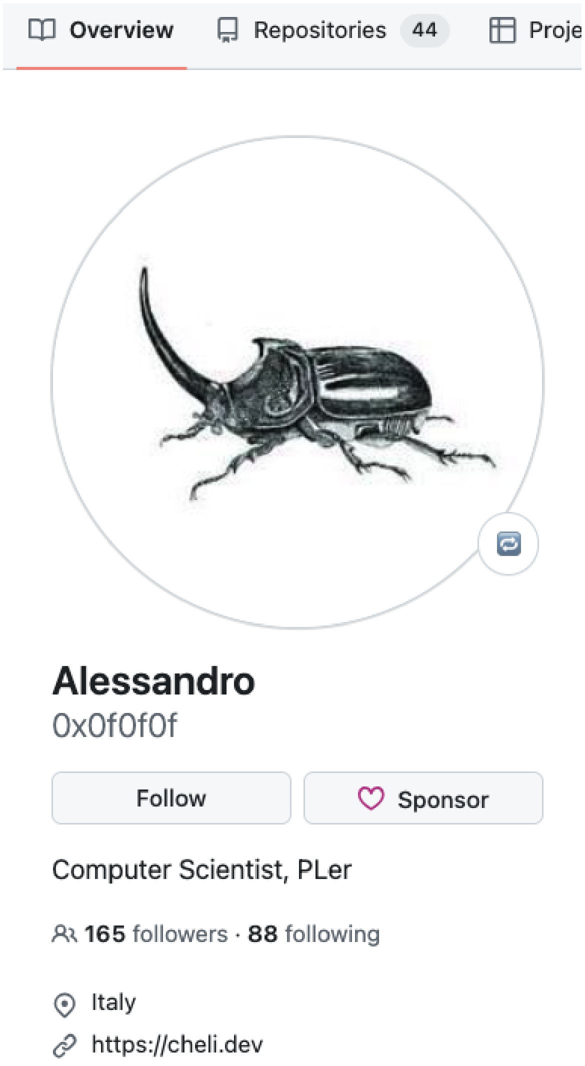
Figure A.1: Time Line



B Data Appendix

B.1 Github User Location

Figure B.1: GitHub User Profile



Alessandro
0x0f0f0f

Computer Scientist, PLer

165 followers · 88 following

Italy
https://cheli.dev

Notes: Profile last accessed on 07 May 2024 at https://github.com/0x0f0f0f.

We scrape, among others, the following key attributes at the user-level for a defined set of locations (location) and span of time-period (i.e. createdAt from 1 Jan 2009 until 11 Apr 2023) from the GitHub GraphQL API:[29]

- login: The user's username (e.g. 0x0f0f0f).
- location: The user's location (e.g. Italy).
- followers: The number of followers the user has (e.g. 165).
- following: The numbers of users the user is following (e.g. 88).
- repositories: A list of user's (public) repositories (e.g. 44).
- createdAt: The date and time the user's account was created.

For more details on the scraping procedure (e.g. the set of locations) and the python implementation please refer to the public GitHub repository https://GitHub.com/sodalabsio/GitHub_scrape.

[29] For more details on the user-level attributes please refer to the official GitHub documentation: https://docs.github.com/en/graphql/reference/objects#user.

B.2 Definition and Construction of Outcome Variables

Table B.1: Variable Definitions

Variable	Description
A – Output Quantity	
Events	Sum of all 15 GitHub event types, i.e., CommitCommentEvent, CreateEvent, DeleteEvent, ForkEvent, GollumEvent, IssuesEvent, IssueCommentEvent, MemberEvent, PublicEvent, PullRequestEvent, PullRequestReviewEvent, PullRequestReviewCommentEvent, PushEvent, ReleaseEvent, WatchEvent
Output	# Commits [PushEvent$.size] + # Issues closed [IssuesEvent$.action == 'closed']+ # Pull requests closed [PullRequestEvent$.action == 'closed'] + # Releases [ReleaseEvent]
Outputs (Shen, 2023)	# Commits [PushEvent$.size] + # Pull requests [PullRequestEvent]
Outputs (Holub and Thies, 2023)	# Commits [PushEvent$.size] + # Comments on issues [IssueCommentEvent]+ # Comments on pull requests [PullRequestReviewCommentEvent] + # Comments on commits [CommitCommentEvent] + # Pull requests [PullRequestEvent] + # Issues [IssuesEvent]
Issues closed	# Issues closed [IssuesEvent$.action == 'closed']
Commits	# Commits [PushEvent$.size]
B – Output Quality	
Pull Requests (PRs) merged	# Closed pull requests that were merged [PullRequestEvent$.pull_request.merged == 'true' & PullRequestEvent$.action == 'closed']
PR merge ratio	(# Closed pull requests that were merged [PullRequestEvent$.pull_request.merged == 'true' & PullRequestEvent$.action == 'closed']) / (# Pull Requests closed [PullRequestEvent$.action == 'closed'])

Table B.1: Variable Definitions *(continued)*

Variable	Description
PR merge ratio (Holub and Thies, 2023)	(# Opened pull requests that were merged [`PullRequest$.pull_request.merged == 'true' & PullRequestEvent$.action == 'closed' & PullRequestEvent$.action == 'opened'`]) / (# Pull Requests opened [`PullRequestEvent$.action == 'opened'`])
C – Task Choice and Complexity	
Pull Requests (PRs) opened	# Pull Requests opened [`PullRequestEvent$.action == 'opened'`]
Avg. files edited per merged PR	Average # files edited per merged pull request [`AVG(PullRequestEvent$.pull_request.changed_files) IF PullRequestEvent$.pull_request.merged == 'true' & PullRequestEvent$.action == 'closed'`]
Avg. Lines added per merged PR	Average # lines added per merged pull request [`AVG(PullRequestEvent$.pull_request.additions) IF PullRequestEvent$.pull_request.merged == 'true' & PullRequestEvent$.action == 'closed'`]
Avg. Lines added per opened PR	Average # lines added per opened pull request [`AVG(PullRequestEvent$.pull_request.additions) IF PullRequestEvent$.action == 'opened'`]
Easy issue closed	# Easy issues closed [`IssuesEvent$.action == 'closed'`]
Interactive Activity	# Comments on issues [`IssueCommentEvent`]+ # Comments on pull requests [`PullRequestReviewCommentEvent`] + # Comments on commits [`CommitCommentEvent`]
D – User Activity	
"Work" hours	Time difference (in hours) between first and last activity (any or "prodctive")]
First activity	Clock Hour of first activity (any or "prodctive")
Last activity	Clock Hour of last activity (any or "prodctive")
Unusual activity (10th %tile)	Dummy = 1 if activity (any or "prodctive") before 6am (10th percentile of *First activity*) or after 9pm (90th percentile of *Last activity*)

Table B.1: Variable Definitions *(continued)*

Variable	Description
Unusual activity (25th %tile)	Dummy = 1 if activity (any or "prodctive") before 8am (25th percentile of *First activity*) or after 7pm (75th percentile of *Last activity*)

Notes: The first column presents the variable name, and the second column provides a detailed description of how each variable is defined. The SQL Google BigQuery code to retrieve the required data is presented in brackets. The keywords to define an issue as "easy" are `good first issues`, `good first bug`, `good-first`, `documentation`, `polish`, `cleanup`, `simple`, `easy`, `small`, `trivial`, `minor help wanted`, `junior job`, `newcomer`, `starter`, `beginner`, `newbie`, `novice`, `low hanging`, `low-hanging` (cf. Holub and Thies, 2023). "Productive" user activity comprises the following events: `PullRequestEvent`, `PullRequestReviewEvent`, `PullRequestReviewCommentEvent`, `PushEvent`, `ReleaseEvent`, `CreateEvent`, `IssueEvent`.

C GitHub User-Level Data

C.1 Descriptive Statistics

Table C.1: Descriptive Statistics

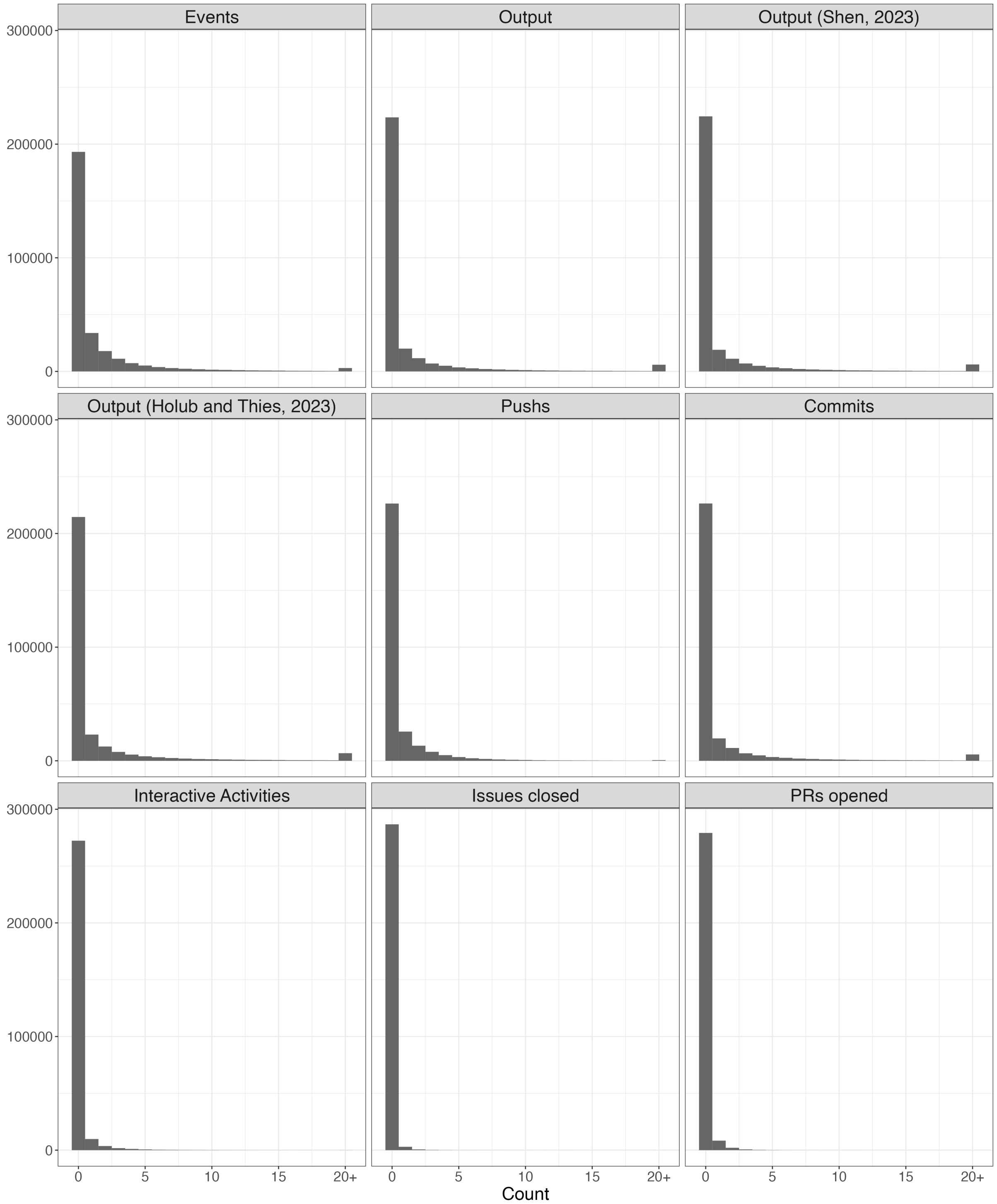
	Overall (N = 36,358)		Less Experienced (N = 18,710)		Experienced (N = 17,648)		Pkg. contributor (N = 5,916)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A - User-Day–Level (N = 290,864)								
Output	0.2314	0.4217	0.2339	0.4233	0.2287	0.4200	0.2688	0.4433
Output (Shen, 2023)	0.2283	0.4198	0.2316	0.4218	0.2249	0.4175	0.2644	0.4410
Output (Holub and Thies, 2023)	0.2624	0.4399	0.2519	0.4341	0.2734	0.4457	0.3312	0.4706
Issue closed	0.0142	0.1183	0.0089	0.0938	0.0198	0.1394	0.0267	0.1611
PR merge ratio	0.0359	0.1846	0.0258	0.1572	0.0467	0.2092	0.0615	0.2381
PR opened	0.0402	0.1964	0.0315	0.1746	0.0494	0.2168	0.0661	0.2484
Avg. lines added per opened PR	0.1581	0.9003	0.1358	0.8648	0.1817	0.9359	0.2356	1.0474
Avg. lines added per merged PR	0.1439	0.8563	0.1148	0.7928	0.1749	0.9179	0.2210	1.0124
Easy issue closed	0.0007	0.0263	0.0005	0.0219	0.0009	0.0302	0.0014	0.0367
Interactive Activity	0.0637	0.2442	0.0340	0.1811	0.0953	0.2936	0.1393	0.3463

B - User–Level (N = 36,358)

	Mean	SD	Min	Median	Max
User creation year	2016.55	3.77	2009	2017	2023
Experienced	0.49	0.50	0	0	1
Pkg. contributions	49.81	447.12	0	0	19638
Pkg. owner	0.05	0.22	0	0	1
Followers	29.49	203.22	0	6	17421
Following	19.67	185.41	0	5	28300
Repositories	29.90	54.53	0	17	3900
Total events	11.93	18.87	1	5	140

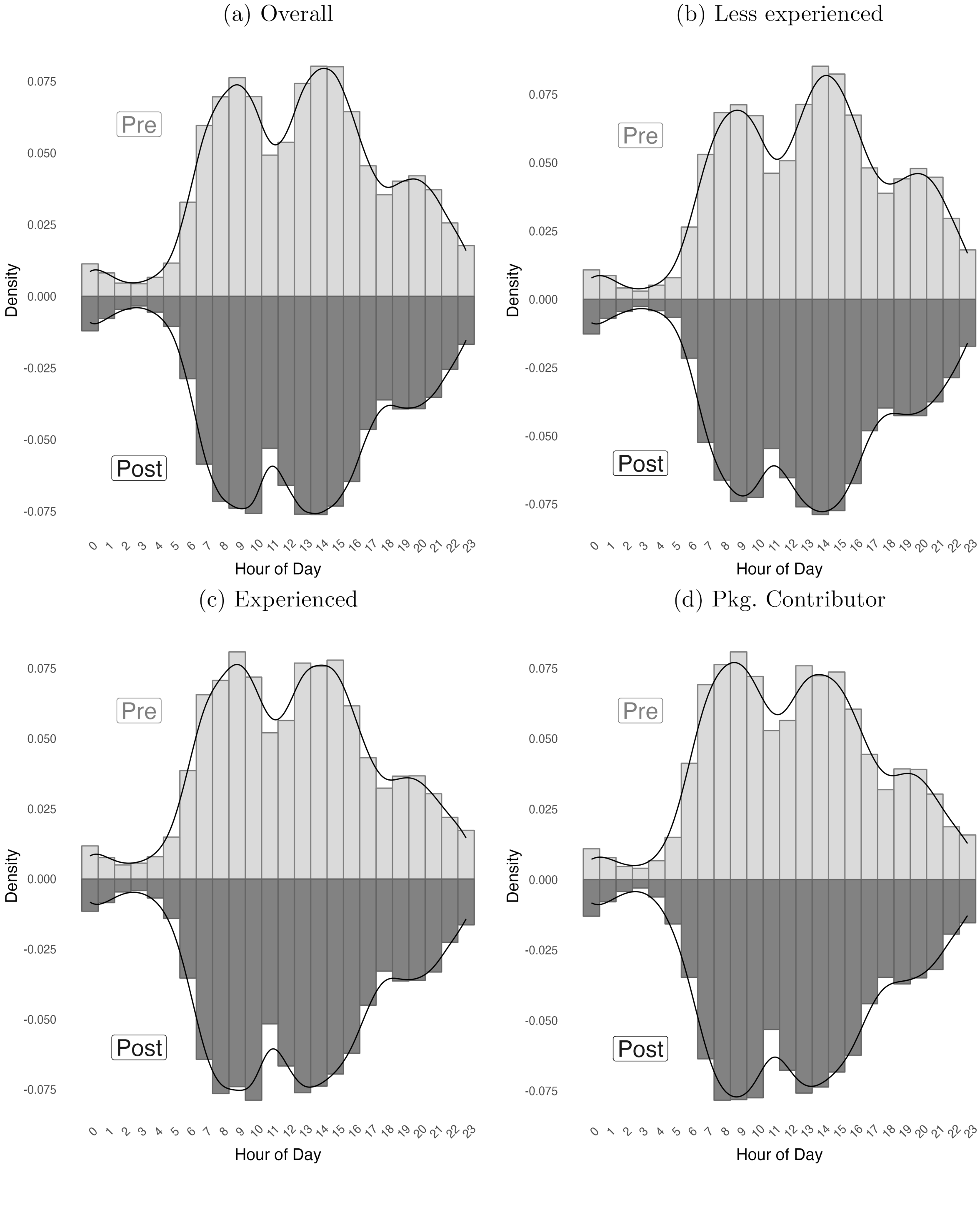
Notes: Panel A presents descriptive statistics for the baseline sample period Pre 27-30.03 – Post 03-06.04. The "Less experienced" sample includes all GitHub user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of unique GitHub user accounts for the entire baseline sample ("Overall") and each of the subsamples is presented in parentheses below. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Panel B provides information on the individual characteristics of all GitHub user accounts in the baseline sample.

Figure C.1: Distribution of Output Quantities



Notes: Daily counts of each action type at the user level for the sample period of March 27–30 (Pre) – April 3–6 (Post) are presented. Counts above 20 are binned and labelled 20+.

Figure C.2: Distribution of Activity across Hours of the Day in Italy Pre-/Post-ChatGPT–Ban


(a) Overall

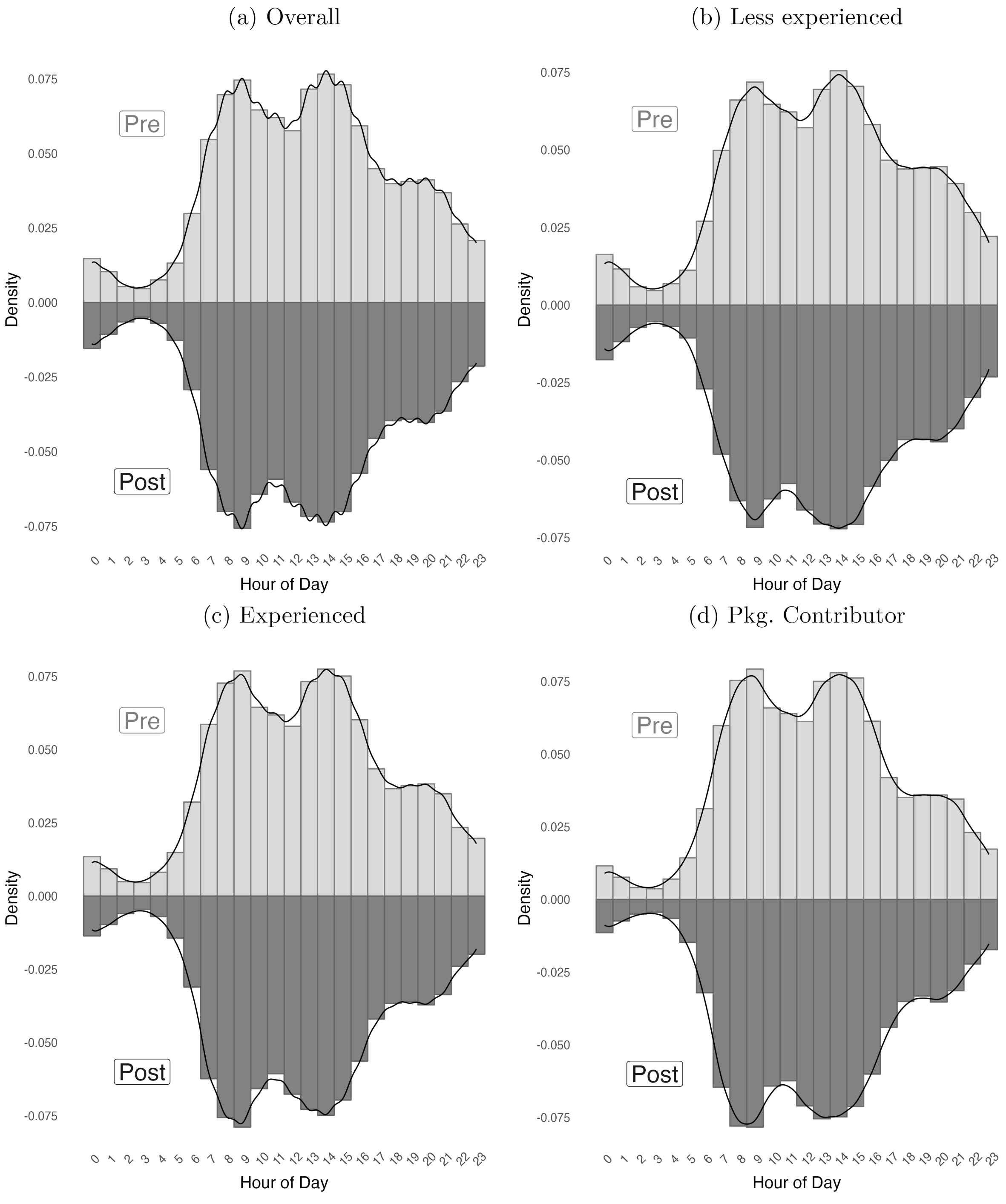
(b) Less experienced

(c) Experienced

(d) Pkg. Contributor

Notes: The distribution of (any) activity by GitHub users across hours of the day is presented separately for the four days pre- and post-ban.

Figure C.3: Distribution of Activity across Hours of the Day in Control Group Pre-/Post-ChatGPT–Ban


(a) Overall

(b) Less experienced

(c) Experienced

(d) Pkg. Contributor

Notes: The distribution of (any) activity by GitHub users across hours of the day is presented separately for the four days pre- and post-ban.

C.2 Additional Results

Table C.2: DID Specification without Linear Time-Trends

		(1)	(2)	(3)	(4)	(5)
A: Output Quantity and Quality						
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023	Issue closed	PR merge ratio
Overall	Treated × Post	0.0116***	0.0128***	0.0147***	-0.0005	0.0056***
(N = 290,864)		(0.0043)	(0.0042)	(0.0045)	(0.0011)	(0.0015)
	Dep. var. mean	0.2314	0.2283	0.2624	0.0142	0.0359
Less experienced	Treated × Post	0.0177***	0.0177***	0.0186***	0.0027**	0.0095***
(N = 149,680)		(0.0061)	(0.0061)	(0.0062)	(0.0013)	(0.0018)
	Dep. var. mean	0.2339	0.2316	0.2519	0.0089	0.0258
Experienced	Treated × Post	0.0052	0.0079	0.0108*	-0.0043**	0.0013
(N = 141,184)		(0.0059)	(0.0059)	(0.0064)	(0.0020)	(0.0025)
	Dep. var. mean	0.2287	0.2249	0.2734	0.0198	0.0467
Pkg. contributor	Treated × Post	0.0148	0.0150	0.0194*	-0.0095**	-0.0001
(N = 47,328)		(0.0109)	(0.0108)	(0.0115)	(0.0040)	(0.0049)
	Dep. var. mean	0.2688	0.2644	0.3312	0.0267	0.0615
B: Task Choice and Complexity						
		PR opened	Avg. lines added per opened PR	Avg. lines added per merged PR	Easy issue closed	Interactive Activity
Overall	Treated × Post	0.0035**	0.0180**	0.0249***	0.0002	0.0011
		(0.0017)	(0.0078)	(0.0070)	(0.0002)	(0.0022)
(N = 290,864)	Dep. var. mean	0.0402	0.1581	0.1439	0.0007	0.0637
Less experienced	Treated × Post	0.0056***	0.0322***	0.0402***	0.0002	0.0016
(N = 149,680)		(0.0021)	(0.0100)	(0.0087)	(0.0002)	(0.0024)
	Dep. var. mean	0.0315	0.1358	0.1148	0.0005	0.0340
Experienced	Treated × Post	0.0012	0.0025	0.0079	0.0001	0.0004
(N = 141,184)		(0.0029)	(0.0121)	(0.0112)	(0.0004)	(0.0039)
	Dep. var. mean	0.0494	0.1817	0.1749	0.0009	0.0953
Pkg. contributor	Treated × Post	-0.0083	-0.0269	-0.0006	-0.0010*	0.0005
(N = 47,328)		(0.0057)	(0.0240)	(0.0219)	(0.0006)	(0.0081)
	Dep. var. mean	0.0661	0.2356	0.2210	0.0014	0.1393

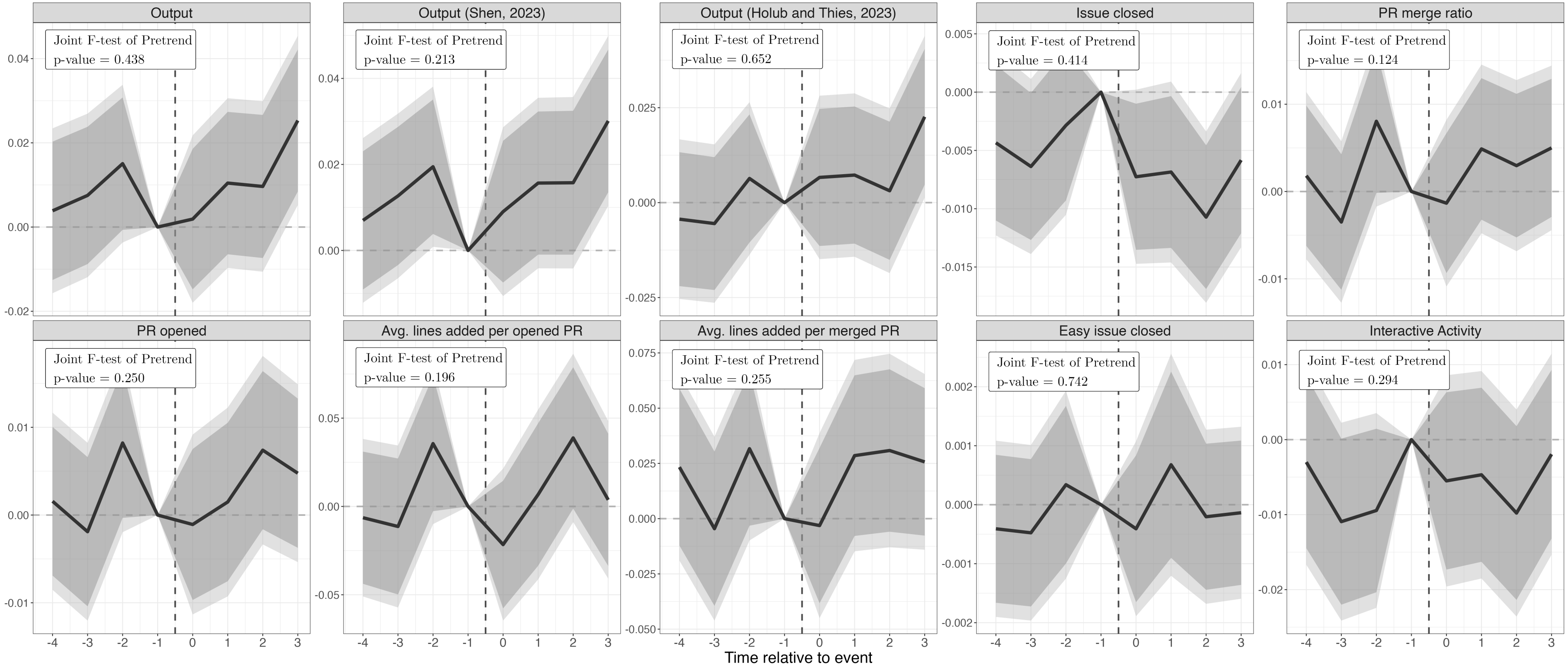
Notes: All specifications include user–fixed effects and day-of-the-week–fixed effects. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Robust standard errors in parentheses are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.3: DID Specification with Country-Specific Linear Time-Trends

		(1)	(2)	(3)	(4)	(5)
A: Output Quantity and Quality						
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023	Issue closed	PR merge ratio
Overall (N = 290,864)	Treated × Post	0.0062 (0.0076)	0.0102 (0.0075)	0.0090 (0.0080)	-0.0045** (0.0022)	0.0045 (0.0030)
	Dep. var. mean	0.2314	0.2283	0.2624	0.0142	0.0359
Less experienced (N = 149,680)	Treated × Post	0.0193* (0.0107)	0.0216** (0.0107)	0.0203* (0.0111)	-0.0017 (0.0023)	0.0118*** (0.0035)
	Dep. var. mean	0.2339	0.2316	0.2519	0.0089	0.0258
Experienced (N = 141,184)	Treated × Post	-0.0079 (0.0107)	-0.0020 (0.0106)	-0.0030 (0.0116)	-0.0077* (0.0039)	-0.0034 (0.0052)
	Dep. var. mean	0.2287	0.2249	0.2734	0.0198	0.0467
Pkg. contributor (N = 47,328)	Treated × Post	0.0181 (0.0205)	0.0164 (0.0200)	0.0270 (0.0215)	-0.0151* (0.0084)	0.0113 (0.0104)
	Dep. var. mean	0.2688	0.2644	0.3312	0.0267	0.0615
B: Task Choice and Complexity						
		PR opened	Avg. lines added per opened PR	Avg. lines added per merged PR	Easy issue closed	Interactive Activity
Overall (N = 290,864)	Treated × Post	0.0000 (0.0034)	-0.0039 (0.0148)	0.0220 (0.0135)	-0.0004 (0.0004)	0.0006 (0.0043)
	Dep. var. mean	0.0402	0.1581	0.1439	0.0007	0.0637
Less experienced (N = 149,680)	Treated × Post	0.0040 (0.0040)	0.0211 (0.0186)	0.0456*** (0.0164)	-0.0004 (0.0004)	0.0037 (0.0045)
	Dep. var. mean	0.0315	0.1358	0.1148	0.0005	0.0340
Experienced (N = 141,184)	Treated × Post	-0.0046 (0.0056)	-0.0324 (0.0237)	-0.0031 (0.0221)	-0.0003 (0.0008)	-0.0028 (0.0077)
	Dep. var. mean	0.0494	0.1817	0.1749	0.0009	0.0953
Pkg. contributor (N = 47,328)	Treated × Post	-0.0221* (0.0117)	-0.0933* (0.0477)	0.0226 (0.0422)	-0.0021* (0.0011)	0.0127 (0.0158)
	Dep. var. mean	0.0661	0.2356	0.2210	0.0014	0.1393

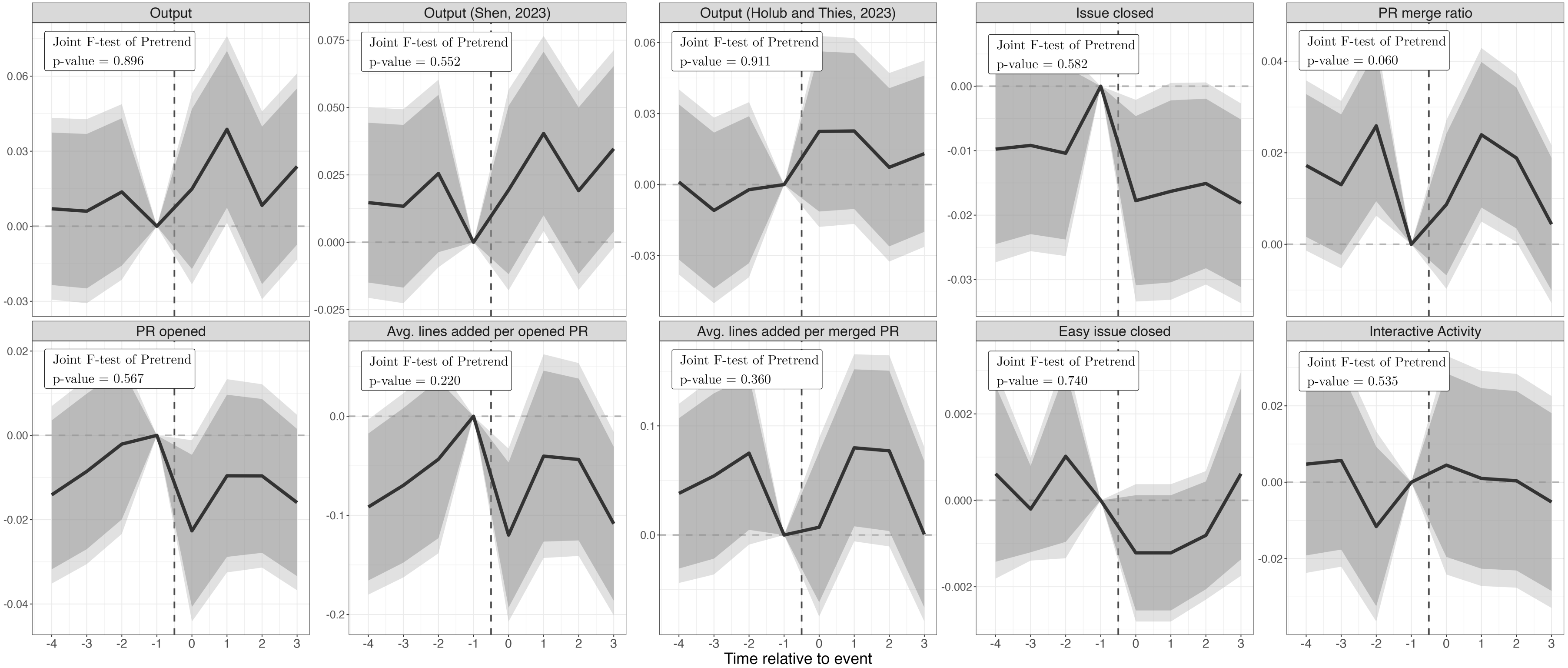
Notes: All specifications include user–fixed effects, day-of-the-week–fixed effects and country-specific linear time trends. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Robust standard errors in parentheses are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure C.4: Experienced Users – Event-Study Estimates



Notes: Event-study estimates across outcomes for "experienced" GitHub user accounts (created before 2017). The sample period spans March 27–30 (*Pre*) and April 3–6 (*Post*). All specifications include user, time, and day-of-the-week fixed effects. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). 95% (90%) confidence intervals for robust standard errors clustered at the user level are depicted in light (dark) grey.

Figure C.5: Pkg. Contributors – Event-Study Estimates



Notes: Event-study estimates across outcomes for "pkg. contributor" GitHub user accounts (owner and/or contributor to a programming package repository). The sample period spans March 27–30 (*Pre*) and April 3–6 (*Post*). All specifications include user, time, and day-of-the-week fixed effects. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). 95% (90%) confidence intervals for robust standard errors clustered at the user level are depicted in light (dark) grey.

Table C.4: Alternative Outcomes

		(1)	(2)	(3)	(4)	(5)
A: Additional Outcomes						
		Any Event	Commit	PR merged	PR merge ratio (Holub & Thies, 2023)	Avg. files edited per merged PR
Overall	Treated × Post	0.0091	0.0084	0.0047	0.0017	0.0119**
(N = 290,864)		(0.0090)	(0.0075)	(0.0031)	(0.0018)	(0.0060)
	Dep. var. mean	0.3358	0.2215	0.0371	0.0130	0.0630
Less experienced	Treated × Post	0.0282**	0.0220**	0.0118***	0.0037	0.0224***
(N = 149,680)		(0.0124)	(0.0106)	(0.0035)	(0.0024)	(0.0074)
	Dep. var. mean	0.3191	0.2272	0.0265	0.0122	0.0498
Experienced	Treated × Post	-0.0116	-0.0061	-0.0029	-0.0004	0.0005
(N = 141,184)		(0.0133)	(0.0105)	(0.0053)	(0.0028)	(0.0098)
	Dep. var. mean	0.3534	0.2155	0.0483	0.0138	0.0769
Pkg. contributor	Treated × Post	0.0311	0.0242	0.0107	-0.0024	0.0277
(N = 47,328)		(0.0233)	(0.0199)	(0.0107)	(0.0050)	(0.0195)
	Dep. var. mean	0.4088	0.2521	0.0635	0.0171	0.0971
B: Intensive Margin						
		Outputs	Outputs (Shen, 2023)	Outputs (Holub & Thies, 2023)	PRs merged	PRs opened
Overall	Treated × Post	0.0122	0.0164	0.0180	0.0053*	0.0010
(N = 290,864)		(0.0145)	(0.0146)	(0.0149)	(0.0032)	(0.0030)
	Dep. var. mean	0.3706	0.3727	0.4215	0.0347	0.0345
Less experienced	Treated × Post	0.0407**	0.0464**	0.0478**	0.0144***	0.0041
(N = 149,680)		(0.0190)	(0.0191)	(0.0193)	(0.0036)	(0.0035)
	Dep. var. mean	0.3538	0.3545	0.3830	0.0247	0.0276
Experienced	Treated × Post	-0.0191	-0.0164	-0.0144	-0.0045	-0.0025
(N = 141,184)		(0.0222)	(0.0224)	(0.0231)	(0.0053)	(0.0049)
	Dep. var. mean	0.3884	0.3919	0.4624	0.0454	0.0418
Pkg. contributor	Treated × Post	-0.0002	0.0033	0.0313	0.0108	-0.0166*
(N = 47,328)		(0.0445)	(0.0445)	(0.0454)	(0.0108)	(0.0101)
	Dep. var. mean	0.4802	0.4852	0.5865	0.0589	0.0556

Notes: All specifications include user–fixed effects, day-of-the-week–fixed effects, and a linear time trend for control and treatment group. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. files edited per PR* and all intensive margin outcomes. Robust standard errors in parentheses are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.5: Placebo Effect in Calendar Week prior to ChatGPT Ban

		(1)	(2)	(3)	(4)	(5)
A: Output Quantity and Quality						
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023)	Issue closed	PR merge ratio
Overall (N = 290,864)	Treated × Post	-0.0010 (0.0070)	-0.0014 (0.0069)	-0.0052 (0.0072)	-0.0016 (0.0022)	-0.0070** (0.0029)
	Dep. var. mean	0.2164	0.2138	0.2444	0.0136	0.0347
Less experienced (N = 149,680)	Treated × Post	-0.0031 (0.0097)	-0.0028 (0.0096)	-0.0038 (0.0099)	-0.0021 (0.0023)	-0.0061* (0.0033)
	Dep. var. mean	0.2163	0.2144	0.2323	0.0083	0.0249
Experienced (N = 141,184)	Treated × Post	0.0004 (0.0100)	-0.0010 (0.0099)	-0.0078 (0.0106)	-0.0010 (0.0038)	-0.0076 (0.0049)
	Dep. var. mean	0.2164	0.2131	0.2573	0.0192	0.0452
Pkg. contributor (N = 47,328)	Treated × Post	0.0096 (0.0189)	0.0094 (0.0185)	0.0000 (0.0198)	-0.0012 (0.0082)	0.0020 (0.0099)
	Dep. var. mean	0.2627	0.2582	0.3208	0.0271	0.0604
B: Task Choice and Complexity						
		PR opened	Avg. lines added per opened PR	Avg. lines added per merged PR	Easy issue closed	Interactive Activity
Overall (N = 290,864)	Treated × Post	-0.0043 (0.0032)	-0.0226 (0.0142)	-0.0206 (0.0131)	-0.0006 (0.0005)	-0.0044 (0.0039)
	Dep. var. mean	0.0381	0.1509	0.1383	0.0006	0.0606
Less experienced (N = 149,680)	Treated × Post	-0.0069* (0.0037)	-0.0246 (0.0178)	-0.0308* (0.0157)	-0.0006 (0.0005)	-0.0027 (0.0042)
	Dep. var. mean	0.0300	0.1285	0.1087	0.0004	0.0315
Experienced (N = 141,184)	Treated × Post	-0.0015 (0.0053)	-0.0213 (0.0228)	-0.0079 (0.0217)	-0.0007 (0.0009)	-0.0062 (0.0069)
	Dep. var. mean	0.0467	0.1747	0.1696	0.0009	0.0913
Pkg. contributor (N = 47,328)	Treated × Post	-0.0018 (0.0107)	-0.0397 (0.0438)	0.0249 (0.0427)	-0.0021 (0.0018)	0.0002 (0.0149)
	Dep. var. mean	0.0656	0.2366	0.2193	0.0012	0.1376

Notes: The *placebo* post-treatment period ranges from Mon 27 March 2023 to Thu 30 March 2023; the corresponding pre-treatment period is from Mon 20 March 2023 to Thu 23 March 2023. All specifications include user–fixed effects, day-of-the-week–fixed effects, and a linear time trend for control and treatment group. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Robust standard errors in parentheses are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.6: Placebo Effect in Calendar Week prior to Easter 2022

		(1)	(2)	(3)	(4)	(5)
A: Output Quantity and Quality						
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023	Issue closed	PR merge ratio
Overall	Treated × Post	0.0109	0.0100	0.0083	-0.0008	0.0034
(N = 145,480)		(0.0110)	(0.0109)	(0.0115)	(0.0037)	(0.0049)
	Dep. var. mean	0.2492	0.2455	0.2872	0.0196	0.0424
Less experienced	Treated × Post	0.0159	0.0152	0.0152	-0.0051	0.0054
(N = 58,400)		(0.0174)	(0.0174)	(0.0179)	(0.0045)	(0.0061)
	Dep. var. mean	0.2448	0.2424	0.2669	0.0114	0.0274
Experienced	Treated × Post	0.0079	0.0069	0.0037	0.0024	0.0023
(N = 87,080)		(0.0142)	(0.0141)	(0.0149)	(0.0054)	(0.0071)
	Dep. var. mean	0.2522	0.2475	0.3008	0.0252	0.0524
Pkg. contributor	Treated × Post	0.0326	0.0261	0.0225	0.0077	-0.0015
(N = 35,680)		(0.0243)	(0.0240)	(0.0252)	(0.0103)	(0.0127)
	Dep. var. mean	0.2853	0.2800	0.3456	0.0311	0.0609
B: Task Choice and Complexity						
		PR opened	Avg. lines added per opened PR	Avg. lines added per merged PR	Easy issue closed	Interactive Activity
Overall	Treated × Post	0.0021	0.0240	0.0413*	0.0000	-0.0071
(N = 145,480)		(0.0050)	(0.0232)	(0.0214)	(0.0005)	(0.0066)
	Dep. var. mean	0.0446	0.1718	0.1593	0.0008	0.0845
Less experienced	Treated × Post	0.0028	0.0171	0.0354	-0.0003	-0.0037
(N = 58,400)		(0.0067)	(0.0338)	(0.0271)	(0.0005)	(0.0081)
	Dep. var. mean	0.0330	0.1382	0.1117	0.0006	0.0441
Experienced	Treated × Post	0.0018	0.0297	0.0472	0.0002	-0.0095
(N = 87,080)		(0.0070)	(0.0315)	(0.0310)	(0.0008)	(0.0096)
	Dep. var. mean	0.0523	0.1943	0.1913	0.0010	0.1116
Pkg. contributor	Treated × Post	0.0006	0.0170	0.0462	-0.0010	0.0037
(N = 35,680)		(0.0128)	(0.0572)	(0.0493)	(0.0012)	(0.0178)
	Dep. var. mean	0.0665	0.2376	0.2159	0.0014	0.1443

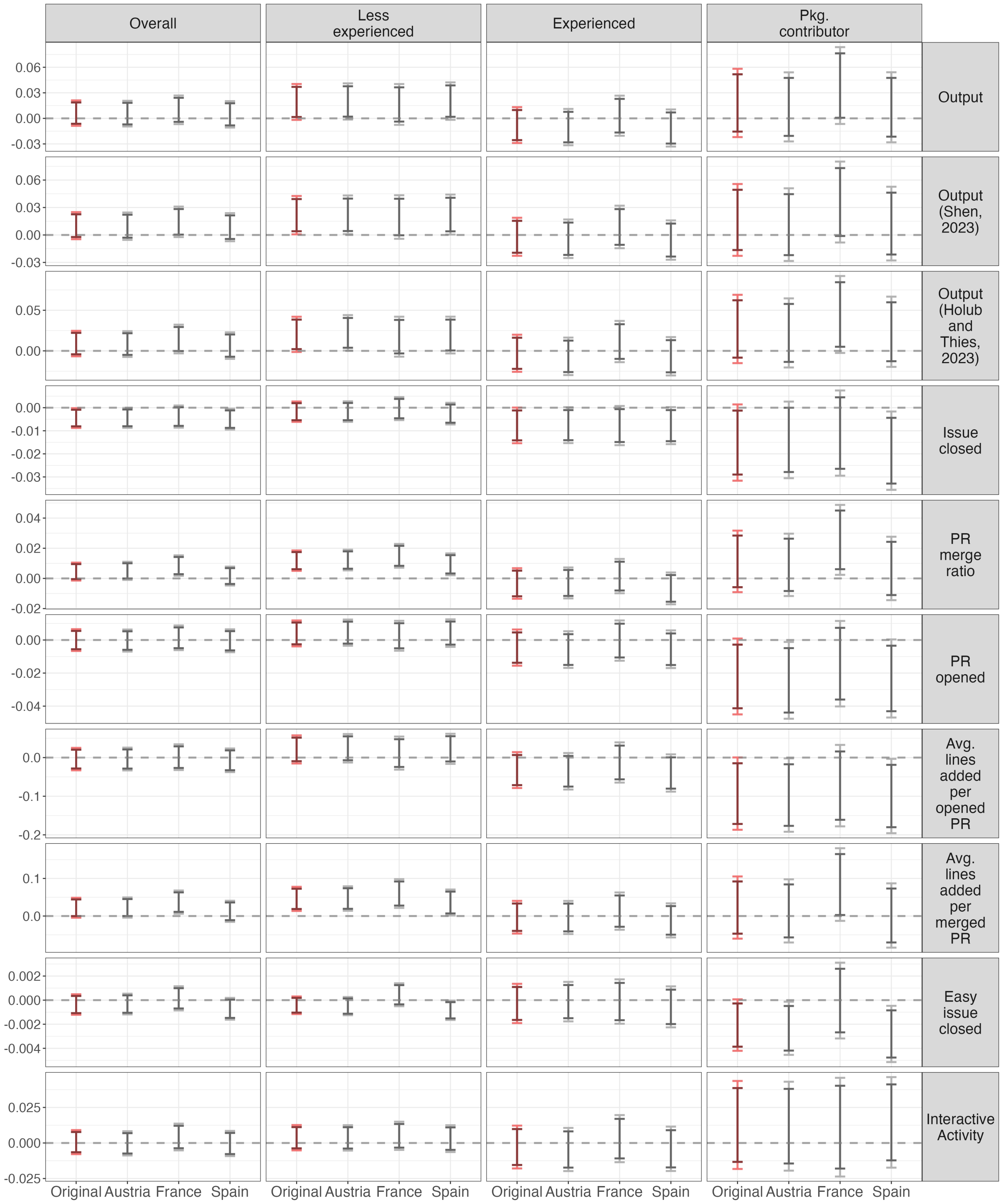
Notes: The *placebo* post-treatment period ranges from Mon 11 April 2022 to Thu 14 April 2022; the corresponding pre-treatment period is from Mon 4 April 2022 to Thu 7 April 2022. All specifications include user–fixed effects, day-of-the-week–fixed effects, and a linear time trend for control and treatment group. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Robust standard errors in parentheses are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.7: Effect of ChatGPT Ban on Working Activity

		(1)	(2)	(3)	(4)	(5)
A: Any Activity						
		“Work” hours	First activity	Last activity	Unusual activity (10th %tile)	Unusual activity (25th %tile)
Overall	Treated × Post	0.0356	-0.2048	-0.1933	0.0032	0.0053
(N = 290,864)		(0.0497)	(0.1777)	(0.1784)	(0.0038)	(0.0063)
	Dep. var. mean	0.9620	11.3925	14.5319	0.0494	0.1292
Less experienced	Treated × Post	0.1344**	-0.2833	-0.0351	0.0083	0.0130
(N = 149,680)		(0.0643)	(0.2543)	(0.2506)	(0.0052)	(0.0084)
	Dep. var. mean	0.8479	11.7579	14.6761	0.0477	0.1196
Experienced	Treated × Post	-0.0742	-0.1427	-0.3592	-0.0026	-0.0029
(N = 141,184)		(0.0772)	(0.2489)	(0.2543)	(0.0055)	(0.0094)
	Dep. var. mean	1.0829	11.0537	14.3982	0.0512	0.1395
Pkg. contributor	Treated × Post	0.1684	-0.0484	-0.1477	0.0069	0.0226
(N = 47,328)		(0.1535)	(0.3616)	(0.3987)	(0.0107)	(0.0180)
	Dep. var. mean	1.4032	10.8529	14.4870	0.0584	0.1652
B: Productive Activities						
		“Work” hours	First activity	Last activity	Unusual activity (10th %tile)	Unusual activity (25th %tile)
Overall	Treated × Post	0.0346	0.0159	0.0083	-0.0004	-0.0010
(N = 290,864)		(0.0428)	(0.2032)	(0.2017)	(0.0032)	(0.0053)
	Dep. var. mean	0.7395	11.4108	14.6531	0.0374	0.0955
Less experienced	Treated × Post	0.1200**	-0.2090	-0.0059	0.0031	0.0082
(N = 149,680)		(0.0567)	(0.2848)	(0.2782)	(0.0045)	(0.0073)
	Dep. var. mean	0.6884	11.7987	14.7975	0.0383	0.0946
Experienced	Treated × Post	-0.0591	0.2347	0.0107	-0.0043	-0.0110
(N = 141,184)		(0.0651)	(0.2906)	(0.2937)	(0.0045)	(0.0077)
	Dep. var. mean	0.7937	11.0049	14.5019	0.0365	0.0965
Pkg. contributor	Treated × Post	0.0822	0.0431	-0.2626	-0.0041	0.0075
(N = 47,328)		(0.1296)	(0.4233)	(0.4664)	(0.0086)	(0.0150)
	Dep. var. mean	0.9935	10.8366	14.4380	0.0416	0.1147

Notes: All specifications include user–fixed effects, day-of-the-week–fixed effects, and a linear time trend for control and treatment group. “Productive activities” comprise: `PullRequestEvent`, `PullRequestReviewEvent`, `PullRequestReviewCommentEvent`, `PushEvent`, `ReleaseEvent`, `CreateEvent`, `IssueEvent`. The number of observations is depicted in parentheses after each sample definition. For *first* and *last activity* only users with at least one activity on the day are included; the number of observations for the “overall”, “less experienced”, “experienced”, and “pkg. contributor” sample in Panel A, respectively, Panel B are 43,868, 40,672, 84,540, 17,802 and 30,717, 32,135, 62,852, 12,663. Robust standard errors in parentheses are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure C.6: “Leave-One-Out” Analysis



Notes: Treatment effect estimates for the baseline DID specification are presented when users from one of the control group countries—i.e. Austria, France or Spain—are consecutively dropped from the sample. Point estimates and 95% (90%) cluster-robust confidence intervals are depicted in light (dark) grey for “Leave-One-Out” estimates and juxtaposed to the “Original” DID estimates (c. Table 1) in light (dark) red.

Table C.8: Effect of ChatGPT Ban accounting for Multiple Hypothesis Testing

		(1)	(2)	(3)	(4)	(5)
A: Output Quantity and Quality						
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023	Issue closed	PR merge ratio
Overall (N = 290,864)	Treated × Post	0.0062 [0.960]	0.0102 [0.602]	0.0090 [0.823]	-0.0045** [0.083]	0.0045 [0.426]
	Dep. var. mean	0.2314	0.2283	0.2624	0.0142	0.0359
Less experienced (N = 149,680)	Treated × Post	0.0193* [0.179]	0.0216** [0.085]	0.0203* [0.157]	-0.0017 [0.960]	0.0118*** [0.000]
	Dep. var. mean	0.2339	0.2316	0.2519	0.0089	0.0258
Experienced (N = 141,184)	Treated × Post	-0.0079 [0.960]	-0.0020 [0.999]	-0.0030 [0.998]	-0.0077* [0.110]	-0.0034 [0.971]
	Dep. var. mean	0.2287	0.2249	0.2734	0.0198	0.0467
Pkg. contributor (N = 47,328)	Treated × Post	0.0181 [0.945]	0.0164 [0.960]	0.0270 [0.710]	-0.0151* [0.179]	0.0113 [0.835]
	Dep. var. mean	0.2688	0.2644	0.3312	0.0267	0.0615
B: Task Choice and Complexity						
		PR opened	Avg. lines added per opened PR	Avg. lines added per merged PR	Easy issue closed	Interactive Activity
Overall (N = 290,864)	Treated × Post	0.0000 [0.999]	-0.0039 [0.998]	0.0220 [0.311]	-0.0004 [0.956]	0.0006 [0.999]
	Dep. var. mean	0.0402	0.1581	0.1439	0.0007	0.0637
Less experienced (N = 149,680)	Treated × Post	0.0040 [0.898]	0.0211 [0.823]	0.0456*** [0.004]	-0.0004 [0.823]	0.0037 [0.960]
	Dep. var. mean	0.0315	0.1358	0.1148	0.0005	0.0340
Experienced (N = 141,184)	Treated × Post	-0.0046 [0.960]	-0.0324 [0.581]	-0.0031 [0.999]	-0.0003 [0.996]	-0.0028 [0.996]
	Dep. var. mean	0.0494	0.1817	0.1749	0.0009	0.0953
Pkg. contributor (N = 47,328)	Treated × Post	-0.0221* [0.132]	-0.0933* [0.109]	0.0226 [0.989]	-0.0021* [0.126]	0.0127 [0.960]
	Dep. var. mean	0.0661	0.2356	0.2210	0.0014	0.1393

Notes: All specifications include user–fixed effects, day-of-the-week–fixed effects, and a linear time trend for the control and treatment group. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). The Romano–Wolf p-values adjusted for multiple hypothesis testing are presented in brackets and calculated with the resampled null distribution from 1000 bootstrap samples with the *Stata* command `rwolf` (Clarke et al., 2020). Robust standard errors are clustered on the user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

D GitHub Repository–User-Level Data

Table D.1: Descriptive Statistics

	Overall		Less experienced		Experienced		Pkg. contributor	
	(N = 4,627 × 11,938)		(N = 3,315 × 4,566)		(N = 4,294 × 7,372)		(N = 3,006 × 2,902)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
A - Repo-User-Day–Level (N = 239,912)								
Output	0.0116	0.1069	0.0080	0.0889	0.0136	0.1158	0.0171	0.1296
Output (Shen, 2023)	0.0108	0.1036	0.0075	0.0860	0.0128	0.1122	0.0163	0.1264
Output (Holub and Thies, 2023)	0.0279	0.1647	0.0192	0.1373	0.0327	0.1780	0.0405	0.1971
Issue closed	0.0024	0.0492	0.0014	0.0376	0.0030	0.0546	0.0034	0.0580
PR merge ratio	0.0030	0.0542	0.0013	0.0358	0.0039	0.0621	0.0047	0.0679
PR opened	0.0054	0.0734	0.0041	0.0641	0.0061	0.0780	0.0080	0.0890
Avg. lines added per opened PR	0.0170	0.2762	0.0134	0.2494	0.0190	0.2902	0.0238	0.3190
Avg. lines added per merged PR	0.0104	0.2167	0.0047	0.1483	0.0135	0.2469	0.0153	0.2541
Easy issue closed	0.0001	0.0094	0.0000	0.0034	0.0001	0.0114	0.0001	0.0089
Interactive Activity	0.0193	0.1377	0.0127	0.1120	0.0230	0.1500	0.0285	0.1663
B - User–Level (N = 11,938)								
	Mean	SD	Min	Median	Max			
User creation year	2015.37	3.51	2009	2015	2023			
Experienced	0.62	0.49	0	1	1			
Pkg. contributions	78.06	563.84	0	0	19514			
Pkg. owner	0.07	0.26	0	0	1			
Followers	49.65	300.24	0	12	17421			
Following	36.54	313.69	0	11	28300			
Repositories	40.40	78.31	0	23	3900			
Total events	2.89	7.47	0	1	98			

Notes: Panel A presents descriptive statistics for the baseline sample period Pre 27-30.03 – Post 03-06.04. The "Less experienced" sample includes all GitHub user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of unique GitHub user accounts for the entire repository × user sample ("Overall") and each of the subsamples is presented in parentheses below. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Panel B provides information on the individual characteristics of all GitHub user accounts in the repository × user sample.

Table D.2: Effect of ChatGPT Ban on GitHub Output at the Repository-Level

		(1)	(2)	(3)	(4)	(5)
A: Output Quantity and Quality						
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023	Issue closed	PR merge ratio
Overall	Treated × Post	-0.0010	-0.0007	0.0029*	-0.0007	-0.0002
(N = 239,912)		(0.0010)	(0.0010)	(0.0016)	(0.0005)	(0.0005)
	Dep. var. mean	0.0116	0.0108	0.0279	0.0024	0.0030
Less experienced	Treated × Post	0.0019	0.0014	0.0047**	0.0007	0.0014***
(N = 86,200)		(0.0015)	(0.0014)	(0.0022)	(0.0006)	(0.0005)
	Dep. var. mean	0.0080	0.0075	0.0192	0.0014	0.0013
Experienced	Treated × Post	-0.0027**	-0.0019	0.0018	-0.0016**	-0.0011
(N = 153,712)		(0.0014)	(0.0013)	(0.0022)	(0.0007)	(0.0007)
	Dep. var. mean	0.0136	0.0128	0.0327	0.0030	0.0039
Pkg. contributor	Treated × Post	-0.0036	-0.0021	0.0011	-0.0025**	0.0001
(N = 63,064)		(0.0025)	(0.0024)	(0.0036)	(0.0012)	(0.0011)
	Dep. var. mean	0.0171	0.0163	0.0405	0.0034	0.0047
B: Task Choice and Complexity						
		PR opened	Avg. lines added per opened PR	Avg. lines added per merged PR	Easy issue closed	Interactive Activity
Overall	Treated × Post	0.0008	0.0022	-0.0008	0.0001	0.0027**
(N = 239,912)		(0.0007)	(0.0027)	(0.0020)	(0.0001)	(0.0013)
	Dep. var. mean	0.0054	0.0170	0.0104	0.0001	0.0193
Less experienced	Treated × Post	0.0019*	0.0059	0.0073***	0.0000	0.0028
(N = 86,200)		(0.0011)	(0.0044)	(0.0024)	(0.0000)	(0.0018)
	Dep. var. mean	0.0041	0.0134	0.0047	0.0000	0.0127
Experienced	Treated × Post	0.0001	0.0002	-0.0058**	0.0001	0.0026
(N = 153,712)		(0.0010)	(0.0034)	(0.0029)	(0.0002)	(0.0018)
	Dep. var. mean	0.0061	0.0190	0.0135	0.0001	0.0230
Pkg. contributor	Treated × Post	0.0003	-0.0030	-0.0015	-0.0002	0.0031
(N = 63,064)		(0.0016)	(0.0058)	(0.0043)	(0.0002)	(0.0031)
	Dep. var. mean	0.0080	0.0238	0.0153	0.0001	0.0285

Notes: All specifications include repository × user fixed effects and day-of-the-week–fixed effects. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Robust standard errors in parentheses are clustered on the repository *times* user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table D.3: Effect of ChatGPT Ban on `PyPI` Repository Contributors' Output

		(1)	(2)	(3)	(4)	(5)
		Output	Output (Shen, 2023)	Output (Holub & Thies, 2023	Issue closed	PR merge ratio
Overall (N = 33,512)	Treated × Post	-0.0042**	-0.0022	-0.0007	-0.0021**	-0.0011
		(0.0021)	(0.0021)	(0.0040)	(0.0009)	(0.0007)
	Dep. var. mean	0.0078	0.0073	0.0227	0.0016	0.0016
Less experienced (N = 12,784)	Treated × Post	-0.0004	0.0017	0.0027	-0.0019*	0.0004
		(0.0029)	(0.0028)	(0.0057)	(0.0010)	(0.0007)
	Dep. var. mean	0.0055	0.0052	0.0173	0.0007	0.0003
Experienced (N = 20,728)	Treated × Post	-0.0069**	-0.0048*	-0.0031	-0.0023*	-0.0021*
		(0.0029)	(0.0029)	(0.0056)	(0.0013)	(0.0011)
	Dep. var. mean	0.0093	0.0086	0.0261	0.0022	0.0023
Pkg. contributor (N = 11,392)	Treated × Post	-0.0060	-0.0024	0.0057	-0.0037**	-0.0025
		(0.0042)	(0.0042)	(0.0082)	(0.0019)	(0.0017)
	Dep. var. mean	0.0121	0.0123	0.0364	0.0011	0.0020

Notes: All specifications include repository × user fixed effects and day-of-the-week–fixed effects. The "Less experienced" sample includes all Github user accounts created after or in 2017 (median), while the "Experienced" sample comprises all GitHub user accounts created before 2017. The "Pkg. contributor" sample comprises all GitHub user accounts that are the owner and/or contributor to a (analytical) programming package repository. The number of observations is depicted in parentheses after each sample definition. A log plus one transformation is applied to *Avg. lines added per PR* (opened or merged). Robust standard errors in parentheses are clustered on the repository *times* user-level: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

E User Adaption to the ChatGPT Ban

Considering that there appears to be mean reversion in the estimated effects toward the end of our sample period, we now turn our attention to adaptation behaviour. The simplest way to circumvent the ChatGPT ban was to use VPN tools or encrypted routing through, for instance, the TOR network.

E.1 Data

We collect daily data on the number of *Google searches* on the topic of "Virtual Private Networks" from *Google Trends* and on the number of users of *TOR*, an open-source software for enabling anonymous communication, from *TOR Metrics* for all 25 countries in the European Union.[30] We retrieve information on both the number of users of "standard" *TOR relays* and of *TOR bridge* relays to examine whether there were changes in the use of, in particular, *TOR bridge* relays, which are not listed publicly and therefore are more difficult for firewalls to identify.[31] We apply a log transformation to both user numbers. The sample period under consideration covers March 13, 2023, the day after the release of ChatGPT-4, until April 7, 2023, the end of the workweek post-ban. Observations on weekends are dropped from the sample since we are interested in the effect of the ban on output. Figure E.1 provides a graphic illustration of the final panel structure.

E.2 Results

To estimate the average treatment effect of the ChatGPT ban on users from Italy, we apply the generalized synthetic control method proposed by Xu (2017). The treatment effect on the treated unit (ATT) is the difference between the actual outcome and its estimated counterfactual. To obtain the counterfactual, a (cross-validated) interactive fixed effects (IFE) model is estimated for the control group data.[32] All IFE models incorporate additive unit and time fixed effects.[33] To draw inference, we rely on the parametric bootstrap procedure suggested by Xu (2017) for settings with a small number of treated units.

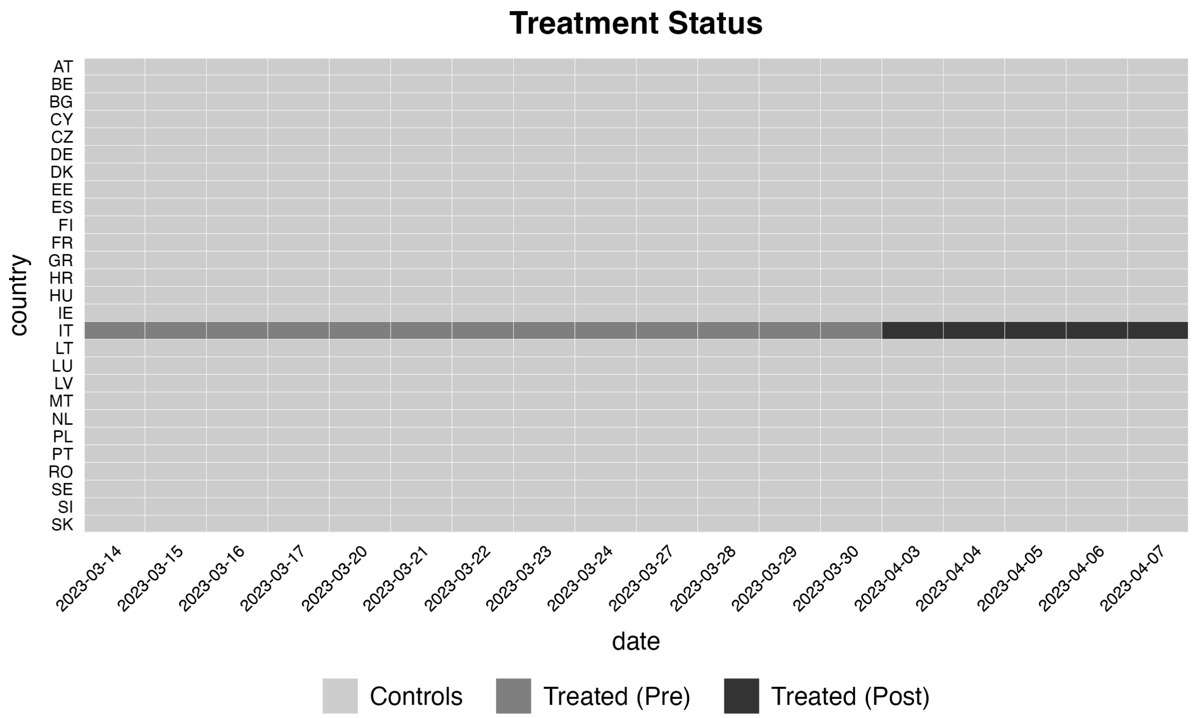
[30] Google trends data have been widely used in economic research as a predictor of human behavioural economic phenomena (Choi and Varian, 2012). For example, Böhme et al. (2020) used Google trends data on migration-related Google search terms to predict international migration, while Ginsberg et al. (2009) used trends data to predict influenza outbreaks.

[31] Note that *TOR bridge* relays can, however, slow down the connection. For more information on bridges vs. "standard" relays, please refer to the official *TOR* documentation at `https://tb-manual.torproject.org/bridges/`.

[32] Specifically, we apply the EM algorithm proposed by Gobillon and Magnac (2016) and implemented in the **R** package `gsynth` (Xu and Liu, 2022), which additionally uses treatment group information for the pre-treatment period, leading to (slightly) more precisely estimated coefficients.

[33] Note that the *Google trends* data are already standardized by country for the selected time period such that we include only time fixed effects in this case when estimating the IFE model.

Figure E.1: Panel Structure of Google Trends and TOR Data

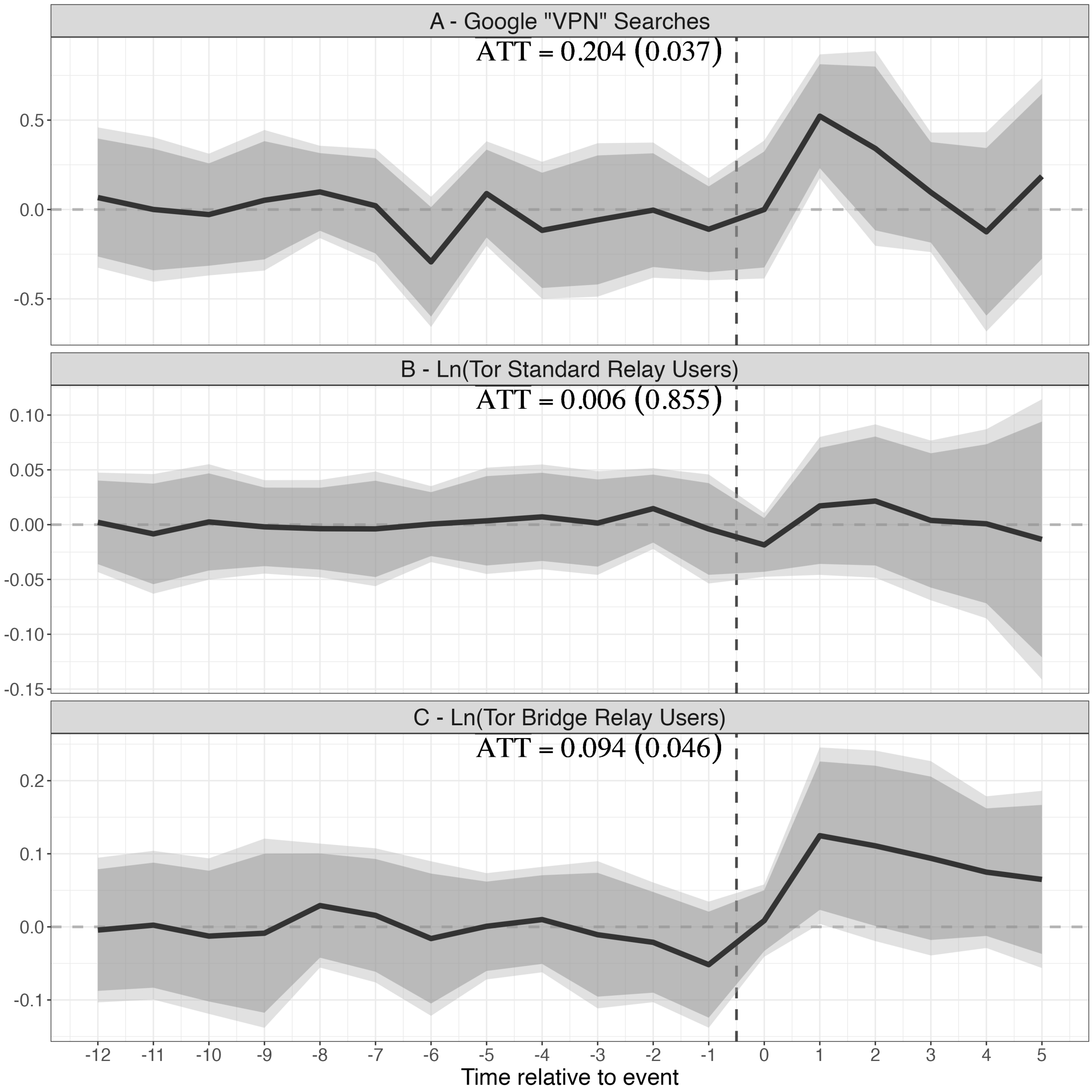


Notes: The panel structure of the datasets used in Section E—i.e., the *(i) Google trends* and the *(ii) TOR* (*"standard"* and *bridge relay*) user datasets—are displayed. Workday dates during the sample period are presented on the x-axis, and *ISO 3166-1 alpha-2* country codes are presented on the y-axis.

The top panel in Figure E.2 presents the effect of the ChatGPT ban on the number of *Google queries* on the topic of VPNs. We observe a significant increase in the share of web searches on this relative to other topics in Italy on the first working day after the ban that slowly vanishes over the next three days. The estimated effect on April 3 is sizeable: the share of searches on VPNs increases by 52.2 percentage points. On average, the share of queries on VPNs was 20.6 percentage points higher in Italy over the workweek. The observed pattern is consistent with Italian users looking for ways to access ChatGPT even after the ban and succeeding after some initial search costs. Our estimates might, however, present only *stated* preferences.

To investigate whether the ban actually led to behavioral changes among Italian users, we look at an alternative outcome: the log number of *TOR* users. The results for *TOR relay* and *TOR bridge* users are presented in Panels B and C of Figure E.2, respectively. While the number of *TOR relay* users shows only a minor increase in the days after the ban, the average treatment effect on the number of Italian *TOR bridge* users is positive and significant on the first workday after the ban. Usage of *TOR bridges* remained elevated for the entire workweek, with an increase in user numbers—on average—of approximately 9.4 percentage points. This pattern is in line with users

Figure E.2: Effect of ChatGPT Ban on Ban Circumvention Technology



Notes: The dynamic treatment effects estimates for the generalized synthetic control method of Xu (2017) are depicted. The top panel presents the ATT for the number of *Google queries* on the topic of VPNs. The bottom panel presents the ATT for *TOR bridge* relay users. The counterfactual for the treated unit (Italy) is estimated with an interactive fixed effects model; 95% (90%) confidence intervals from the parametric bootstrap procedure proposed by Xu (2017) are displayed in light (dark) grey. Additionally, the *mean* ATT over the workweek after the ChatGPT ban and its *p*-value (in parentheses) are presented.

resorting to *bridge* over "standard" relays to minimize the chance of their being denied access to ChatGPT since the former are more difficult for firewalls to identify.[34]

Overall, our findings are consistent with Italian users looking for and finding ways to

[34] For a discussion on denial of ChatGPT access, see the following OpenAI forum discussion: https://community.openai.com/t/access-denied-error-1020/38758/23.

circumvent the blocked access to ChatGPT.

Appendix References

Böhme, Marcus H., André Gröger, and Tobias Stöhr, "Searching for a better life: Predicting international migration with online search keywords," *Journal of Development Economics*, 2020, *142*, 102347.

Choi, Hyunyoung and Hal Varian, "Predicting the present with Google trends," *Economic Record*, 2012, *88* (s1), 2–9.

Clarke, Damian, Joseph P. Romano, and Michael Wolf, "The Romano–Wolf multiple-hypothesis correction in Stata," *Stata Journal*, 2020, *20* (4), 812–843.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, "Detecting influenza epidemics esing search engine query data," *Nature*, 2009, *457*, 1012–1014.

Gobillon, Laurent and Thierry Magnac, "Regional policy evaluation: Interactive Fixed Effects and Synthetic Controls," *Review of Economics and Statistics*, 07 2016, *98* (3), 535–551.

Holub, Felix and Beate Thies, "Air quality, high-skilled worker productivity and adaptation: Evidence from GitHub," CRC TR 224 Discussion Paper Series crctr224_2023_402, University of Bonn and University of Mannheim, Germany March 2023.

Shen, Lucas, "Does working from home work? A natural experiment from lockdowns," *European Economic Review*, 2023, *151*, 104323.

Xu, Yiqing, "Generalized Synthetic Control Method: Causal inference with Interactive Fixed Effects Models," *Political Analysis*, 2017, *25* (1), 57–76.

_ **and Licheng Liu**, *gsynth: Generalized Synthetic Control Method* 2022. R package version 1.2.1.