Brodeur, Abel et al.

Working Paper

Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science

I4R Discussion Paper Series, No. 195

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Brodeur, Abel et al. (2025) : Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science, I4R Discussion Paper Series, No. 195, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/308508

INSTITUTE for REPLICATION

No. 195
I4R DISCUSSION PAPER SERIES

Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science

Abel Brodeur
Alexandru Marcoci
Derek Mikola
Rohan Alexander
Tom Stafford
Gunther Bensch
David Valenta
Juan P. Aparicio
Bruno Barbarioli
Lachlan Deer
Lars Vilhuber
et al.

January 2025

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 195

Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science

Abel Brodeur[1,2], David Valenta[1,2], Alexandru Marcoci[3,4], Juan P. Aparicio[2], Derek Mikola[2], Bruno Barbarioli[2], Rohan Alexander[5], Lachlan Deer[6], Tom Stafford[7], Lars Vilhuber[8], Gunther Bensch[9], et al.[10]

[1]University of Ottawa/Canada
[2]Institute for Replication
[3]University of Nottingham/Great Britain
[4]University of Cambridge/Great Britain
[5]University of Toronto/Canada
[6]Tilburg University/The Netherlands
[7]University of Sheffield/Great Britain
[8]Cornell University, Ithaca/USA
[9]RWI – Leibniz Institute for Economic Research, Essen/Germany
[10]see next pages for full author list

JANUARY 2025

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science

Abel Brodeur[12*], David Valenta[12], Alexandru Marcoci[34], Juan P. Aparicio[2]
Derek Mikola[2], Bruno Barbarioli[2], Rohan Alexander[5], Lachlan Deer[6]
Tom Stafford[7], Lars Vilhuber[8], Gunther Bensch[9], et al.[10]

[1]Department of Economics, University of Ottawa, Ottawa & K1N 9A7, Canada.

[2]Institute for Replication, University of Ottawa, Ottawa & K1N 9A7, Canada.

[3]School of Politics and International Relations, University of Nottingham, Nottingham NG7 2RD, UK.

[4]Centre for the Study of Existential Risk, University of Cambridge, Cambridge CB2 1SB, UK.

[5]Faculty of Information and Department of Statistical Sciences, University of Toronto, Toronto, M5S 3G6, Canada.

[6]Department of Marketing, Tilburg University, Tilburg, 5037AB, The Netherlands.

[7]School of Psychology, University of Sheffield, Sheffield, UK.

[8]Department of Economics, ILR School, Cornell University, Ithaca, NY, 14853, United States.

[9]RWI – Leibniz Institute for Economic Research, Essen, 45128, Germany.

[10]See next page for full author list. [*]Corresponding author. Email: abrodeur@uottawa.ca

This study evaluates the effectiveness of varying levels of human and artificial intelligence (AI) integration in reproducibility assessments of quantitative social science research. We computationally reproduced quantitative results from published articles in the social sciences with 288 researchers, randomly assigned to 103 teams across three groups — human-only teams, AI-assisted teams and teams

whose task was to minimally guide an AI to conduct reproducibility checks (the "AI-led" approach). Findings reveal that when working independently, human teams matched the reproducibility success rates of teams using AI assistance, while both groups substantially outperformed AI-led approaches (with human teams achieving 57 percentage points higher success rates than AI-led teams, $p < 0.001$). Human teams were particularly effective at identifying serious problems in the analysis: they found significantly more major errors compared to both AI-assisted teams (0.7 more errors per team, $p = 0.017$) and AI-led teams (1.1 more errors per team, $p < 0.001$). AI-assisted teams demonstrated an advantage over more automated approaches, detecting 0.4 more major errors per team than AI-led teams ($p = 0.029$), though still significantly fewer than human-only teams. Finally, both human and AI-assisted teams significantly outperformed AI-led approaches in both proposing (25 percentage points difference, $p = 0.017$) and implementing (33 percentage points difference, $p = 0.005$) comprehensive robustness checks. These results underscore both the strengths and limitations of AI assistance in research reproduction and suggest that despite impressive advancements in AI capability, key aspects of the research publication process still require human substantial human involvement.

Abel Brodeur (University of Ottawa; Institute for Replication), David Valenta (University of Ottawa), Alexandru Marcoci (University of Nottingham, University of Cambridge), Juan P. Aparicio (University of Ottawa; Institute for Replication), Derek Mikola (University of Ottawa; Institute for Replication), Bruno Barbarioli (University of Ottawa; Institute for Replication), Rohan Alexander (University of Toronto), Lachlan Deer (Tilburg University), Tom Stafford (Sheffield University), Lars Vilhuber (Cornell University), Gunther Bensch (RWI - Leibniz Institute for Economic Research), Mohamed Abdelhady (Carleton University), Yousra Abdelmoula (Carleton University), Ghina Abdul Baki (University of Ottawa), Tomás Aguirre (Centre for the Governance of AI), Sriraj Aiyer (University of Oxford), Shumi Akhtar (The University of Sydney), Farida Akhtar (Macquarie University), Melle R. Albada (Vienna University of Economics and Business), Micah Altman (MIT), David Angenendt (Technical University of Munich), Zahra Arjmandi Lari (Independent researcher), Jorge Armando De León Tejada (Universidad del Rosario), Igor Asanov (Inter-

national Center for Higher Education Research and Faculty of Economics, University of Kassel), Anastasiya-Mariya Asanov Noha (University of Kassel, INCHER), Rebecca Ashong (University of Ghana), Tobias Auer (London School of Economics), Francisco J. Bahamonde-Birke (Tilburg University), Bradley J. Baker (Temple University), Söhnke M. Bartram (University of Warwick and CEPR), Dongqi Bao (University of Zurich), Lucija Batinovic (Linköping University), Tommaso Batistoni (University of Oxford), Monica Beeder (NHH Norwegian School of Economics), Louis-Philippe Beland (Carleton University), Carsten Bienz (Norwegian School of Economics), Christ Billy Aryanto (Faculty of Psychology, Atma Jaya Catholic University of Indonesia), Cylcia Bolibaugh (University of York), Carl Bonander (University of Gothenburg), Ramiro Bravo (Faculty of Biology, Medicine and Health, The University of Manchester), Katherine Brennan (Bank of Canada), Egor Bronnikov (Maastricht University; George Mason University), Stephan Bruns (Hasselt University), Nino Buliskeria (Nazarbayev University), Sara Caicedo-Silva (Universidad de los Andes), Andrea Calef (University College London, School of Management), Solomon Caulker (United Methodist University Sierra Leone), Simonas Cepenas (ISM University of Management and Economics), Arthur Chatton (Université Laval), Zirou Chen (University of Toronto), Ngozi Chioma Ewurum (Michael Okpara University of Agriculture, Umudike, Nigeria), Anda-Bianca Ciocîrlan (University of Sheffield), Felix J. Clouth (Tilburg University), Jason Collins (University of Technology Sydney), Nikolai Cook (Wilfrid Laurier University), Cesar Cornejo (The London School of Hygiene & Tropical Medicine), João Craveiro (University of Sheffield), Jing Cui (University of Ottawa), Niveditha Chalil Vayalabron (School of Earth and Planetary Science, National Institute of Science Education and Research, India), Christian Czymara (Goethe-Universitaet Frankfurt), Carlos Daniel Bermúdez Jaramillo (Universidad del Rosario), Hannes Datta (Tilburg University), Lien Denoo (Tilburg University), Arshia Dhaliwal (Carleton University), Nency Dhameja (Binghamton University), Elodie Djemai (Université Paris-Dauphine), Erwan Dujeancourt (Stockholm University), Uğurcan Dündar (Vienna University of Economics and Business), Thibaut Duprey (Bank of Canada), Yasmine Eissa (The American University in Cairo), Youssef El Fassi (HEC Lausanne), Ismail El Fassi (University of St. Gallen), Keaton Ellis (UC Berkeley), Ali Elminejad (Nazarbayev University), Mahmoud Elsherif (University of Leicester), Aysil Emirmahmutoglu (NHH Norwegian School of Economics), Giulian Etingin-Frati (University of Zurich), Emeka Eze (Michael Okpara University of Agriculture), Jan Fabian Dollbaum (University College Dublin),

Jan Feld Victoria (University of Wellington), Andres Felipe Rengifo Jaramillo (Business School; Universidad de los Andes), Guidon Fenig (University of Ottawa), Victoria Fernandes (Bank of Canada), Lenka Fiala (University of Bergen), Lukas Fink (FU Berlin), Sara Fish (Harvard University), Jack Fitzgerald (Vrije Universiteit Amsterdam), Rachel Forshaw (Heriot-Watt University), Alexandre Fortier-Chouinard (Université Laval), Louis Fréget (CEPREMAP), Joris Frese (European University Institute), Jacopo Gabani (World Bank; Centre for Health Economics, University of York), Sebastian Gallegos (UAI Business School), Max C. Gamill (University of Sheffield), Attila Gáspár (HUN-REN Centre for Economic and Regional Studies), Romain Gauriot (Deakin University), Evelina Gavrilova (NHH Norwegian School of Economics), Diogo Geraldes (University College Dublin), Giulio Giacomo Cantone (University of Sussex), Grant Gibson (McMaster University & CRDCN), Dirk Goldschmitt (University of Sheffield), Amélie Gourdon-Kanhukamwe (King's College London), Andrea Gregor de Varda (University of Milano-Bicocca), Idaliya Grigoryeva (UC San Diego), Alexi Gugushvili (University of Oslo), Aaron H.A. Fletcher (University of Sheffield), Florian Habermann (University of Lausanne), Márton Hablicsek (Leiden University), Joanne Haddad (Université Libre de Bruxelles), Jonathan D. Hall (University of Alabama), Olle Hammar (Linnaeus University and Institute for Futures Studies), Malek Hassouneh (University of Toronto), Carina I Hausladen (ETH Zürich), Sophie C. F. Hendrikse (Tilburg University), Matthew Hepplewhite (University of Oxford), Anson T. Y. Ho (Toronto Metropolitan University), Senan Hogan-Hennessy (Cornell University), Elliot Howley (University of Nottingham), Gaoyang Huang (Swiss Federal Institute of Technology in Zurich), Héloïse Hulstaert (Hasselt University; Liège University), Zlatomira G. Ilchovska (University of York; University of Birmingham), Niklas Jakobsson (Karlstad University), Joakim Jansson (Linnaeus University; Research Institute of Industrial Economics), Ewa Jarosz (University of Warsaw), Hossein Jebeli (Bank of Canada), Yanchen Jiang (Harvard University), Hiba Junaid (Bart's Life Sciences, Bart's Health NHS Trust; Queen Mary university of London), Rohan Kalluraya (Cornell University), Edmund Kelly (University of Oxford), Eva Kimel (University of York), Sorravich Kingsuwankul (Vrije Universiteit Amsterdam), Valentin Klotzbücher (University of Freiburg), Daniel Krähmer (University of Munich), Pijus Krūminas (ISM University of Management and Economics), Nicholas Kruus (Schelling Research), Essi Kujansuu (University of Innsbruck), Christoph F. Kurz (Ludwig-Maximilians-Universität Munich) Stephan Küster (Freie Universität Berlin), Blake Lee-Whiting (University of Toronto), Felix

Lewandowski (University of Nottingham), Tongzhe Li (University of Guelph), Ruoxi Li (Yale University), Dan Liu (Australian National University), Jiacheng Liu (Purdue University), Helix Lo (University of Tokyo), Katharina Loter (Tilburg University), Felipe Macedo Dias (Cornell University), Christopher R. Madan (University of Nottingham), Nicolas Mäder (University of San Diego), Marco Mandas (University of Cagliari), Jan Marcus (FU Berlin), Diego Marino Fages (Durham University), Xavier Martin (Tilburg University), Ryan McWay (University of Minnesota), Daniel Medina-Gaspar (Universidad EAFIT), Sisi Meng (University of Notre Dame), Lingyu Meng (University of Sheffield), Alex P. Miller (University of Southern California), Thibault Mirabel (Equalis Capital), Dibya Deepta Mishra (Rice University), Sumit Mishra (Krea University), Belay W. Moges (Dilla University), Morteza Mohandes Mojarrad (Tilburg university), Myra Mohnen (University of Ottawa), Louis-Philippe Morin (University of Ottawa), Fabio Motoki (University of East Anglia), Lucija Muehlenbachs (University of Calgary), Gastón Mullin (Tilburg University), Andreea Musulan (University of Montreal), Sara Muzzì (University of Milano Bicocca), James A. C. Myers (University of Sheffield), Florian Neubauer (RWI - Leibniz Institute for Economic Research), Tuan Nguyen (Hasselt University), Ali Niazi (University of Calgary), Ardyn Nordstrom (Carleton University), Bartłomiej Nowak (Cardinal Stefan Wyszynski University), Daneal O'Habib (Bank of Canada), Tim Ölkers (University of Göttingen), Justin Ong (University of Sheffield), Valeria Orozco Castiblanco (IESE, Universidad de Navarra), Ömer Özak (SMU), Ali I. Ozkes (SKEMA Business School, GREDEG, Université Côte d'Azur), Mikael Paaso (Erasmus University Rotterdam), Shubham Pandey (Universität Osnabrück), Varvara Papazoglou (University of Sheffield), Romeo Penheiro (University of Houston), Linh Pham (Lake Forest College), Ulrike Phieler (Vienna University of Economics and Business), Peter Pütz (Bielefeld University), Quan Qi (University at Albany, SUNY), Jingyi Qiu (University of Michigan), David A. Reinstein (The Unjournal), Juuso Repo (INVEST Flagship Research Center, University of Turku), Nicolas Rudolf (University of Lausanne), Shree Saha (Cornell University), Orkun Saka (City, University of London), Chiara Saponaro (University of Milano-Bicocca), Georg Sator (University of Nottingham), Martijn Schoenmakers (Tilburg University), Raffaello Seri (InsIDE Lab, DiEco, Università degli Studi dell'Insubria), Meet Shah (Toronto Metropolitan University), Paul Sibille (University of Liege), Christoph Siemroth (University of Essex), Vladimir Skavysh (Bank of Canada), Ben Slater (University of Cambridge), Wenting Song (Bank of Canada), Stefan Staubli (University of Calgary), Tobias Steindl (Univer-

sity of Regensburg), Nomwendé Steven Waongo (University of Ottawa), Paul Stott (University of Manchester), Stephenson Strobel (McMaster University), Roshini Sudhaharan (Tilburg University), Pu Sun (University of Ottawa), Scott D. Swain (Clemson University), Oleksandr Talavera (University of Birmingham), Hanz M. Tantiangco (University of Sheffield), Georgy Tarasenko (Cornell University), Boyd Tarlinton (Department of Primary Industries, QLD), Mariam Tarraf (Carleton University), Ken Teoh (International Monetary Fund), Rémi Thériault (Université du Québec à Montréal), Bethan Thompson (SRUC), Tonghui Tian (Carleton University), Wenjie Tian (University of Ottawa), Manuel Tobias Rein (Tilburg University), Emmanuel Tolani (University of Bonn), Nicolai Topstad Borgen (University of Oslo), Solveig Topstad Borgen (University of Oslo), Javier Torralba (Tilburg University), Carolina Velez-Ospina (World Bank), Man Wai Mak (Carleton University), Lukas Wallrich (Birkbeck, University of London), Zeyang Wang (Vanderbilt University), Leah Ward (University of Manchester), Matthew D. Webb (Carleton University), Duncan Webb (Princeton University), Bryan S. Weber (College of Staten Island, CUNY), Christoph Weber (ESSCA School of Management), Wei-Chien Weng (National Taiwan University), Christian Westheide (University of Vienna), Tom Wilkinson (University of Sheffield), Kwong-Yu Wong (National University of Singapore), Marcin Wroński (Collegium of World Economy, SGH Warsaw School of Economics), Zhuangchen Wu (University of Birmingham), Qixia Wu (University of Ottawa), Victor Y. Wu (Stanford University), Bohan Xiao (University of Ottawa), Feihong Xu (Northwestern University), Cong Xu (National Chengchi University; Aalto University), Pranav Yadav (Tilburg University), Yu Yang Chou (University College London), Luther Yap (Princeton University), Myra Yazbeck (University of Ottawa), Zuzanna Zagrodzka (University of Sheffield), Tahreen Zahra (Carleton University), Mirela Zaneva (University of Oxford), Xiaomeng Zhang (Nanjing Audit University), Ziwei Zhao (University of Lausanne; Swiss Finance Institute), Han Zhong (University of Toronto), Aras Zirgulis (ISM University of Management and Economics), Jiacheng Zou (Columbia University), Floris Zoutman (NHH Norwegian School of Economics), Christelle Zozoungbo (Penn State University).

1 Introduction

Reproducibility is a cornerstone of robust empirical research. This is particularly so in empirical work where complex methodologies and data handling techniques are common. (*1–10*) Despite advancements in reproducibility protocols, (*11*) concerns persist regarding the accuracy and reliability of published findings. (*12–19*) While the current reproducibility and replicability crisis in the behavioral and social sciences is a complex problem, challenges with peer review, the methodological expertise required to design and then assess quantitative studies, as well as transparency of reporting have been identified as key causes. This study investigates the role AI tools, that is large language models (LLMs) (*20, 21*), could play in supporting human researchers, data editors and scientific journals to computationally reproduce social science findings. Our focus is on three modes of AI and human interaction: human-only teams (the "human" approach), human teams with AI assistance (the "AI-assisted" approach) and teams whose task was to minimally guide an AI to conduct reproducibility checks (the "AI-led" approach). We focus on ChatGPT, powered by GPT-4/4o.

This study tests how effectively AI can reproduce studies and handle the nuances of quantitative social science research, particularly in complex cases where coding errors or methodological inconsistencies arise. We employ a randomized controlled trial design involving three treatment arms to assess the effectiveness of AI in evaluating reproducibility. By examining three approaches, we aim to understand and contribute to a large literature documenting the benefits and limitations of human-AI teams, as well as full automation. (*22–37*) This is crucial for science as current methods for performing computational reproducibility and robustness checks are expensive, time consuming (*38, 39*) and require advanced technical skills, and a growing body of literature documents the potential pitfalls of integrating human and artificial intelligence such as overreliance and expertise erosion. (*40, 41*)

We examine three primary outcomes across the treatment groups: (1) computational reproducibility success rates, (2) error detection capabilities, and (3) robustness check quality. By understanding these facets, this study contributes to broader discussions on AI, offering insights into the optimal balance of human and AI involvement in research reproduction tasks.

2 Procedures

2.1 Events

The first ten coauthors organized seven AI replication games between February and November 2024. All remaining coauthors (graduate students, postdoctoral fellows, professors or researchers from non-academic organizations with a PhD; see Table S3) and a few organizers participated in one of those games. Randomization was carried out in two steps for each of the seven events. In step one, coauthors were randomly assigned to a team of three to evaluate the reproducibility of a quantitative social science article. The randomization in step one was conditional on the software abilities and preferences reported by participants (Stata or R) and the mode of participation (in person or virtual). In step two, each team was randomly assigned to one of three treatment arms: AI-led, AI-assisted, or human.

Each team was assigned a study from a leading behavioral, economics, political science, or psychology journal and tasked to computationally reproduce a few pre-defined numerical results, detect coding errors and data irregularities, and suggest two robustness checks (see Supplementary Materials Study Selection for list of studies). Each event had two studies with known coding errors identified by the lead authors in a previous study, but not publicly released when included in the AI replication game: one whose reproduction package was written in Stata and another that was written in R. Each team was assigned the study that matched their software abilities and preferences in attendance mode. Teams had no information about the study they would be reproducing until the day of the event. In total, 12 studies were used.

Teams were revealed the materials for the event at about 09:00 local-time the day of the event. We did this *via* email, where we shared with them an online Open Science Framework webpage which contained: the journal article and online appendix as PDFs, the original authors' replication package (R or Stata), and screenshots of the exhibit to reproduce from the article (see Supplementary Materials). Of note, the screenshots were implemented after the pilot event as they could be useful to AI-led teams as they encode the information in different way than the PDF files. Teams had seven hours to complete three tasks: (i) computationally reproduce a few pre-determined results, (ii) detect coding errors, and (iii) suggest and implement up to two robustness checks. Teams could leave before the end of the event if they believed they had completed their tasks. Upon completion, teams

were asked to email us a (templated) time log that documents whether they completed computational reproducibility and includes all coding errors uncovered and two robustness checks. AI-assisted and AI-led teams also had to provide their AI conversation history.

Access to a paid subscription of ChatGPT, powered by GPT-4, and later also GPT-4o models, was provided to all coauthors in the AI-assisted and AI-led teams. While other models were available during later events, like the chain of thought using o1-preview, this model was not capable of processing files and therefore was of little use to the AI-led teams. AI-assisted and AI-led teams took part in a mandatory 1-hour long training on the usage of ChatGPT (*42*). The training could be viewed live or as a recording. Human teams were allowed to take part in the training at their discretion. Additional details on our AI training and models available and their capabilities can be found in Supplementary Materials section 3.

Human teams were not allowed to use ChatGPT or any other AI tool. The AI-assisted groups were allowed to use ChatGPT without limitation (but no other AI tool). This group could have chosen not to use ChatGPT at all if they preferred to. Those in the AI-led teams were not allowed to read the article, look at the data or go through the analysis code. AI-led teams had to conduct their analysis only through ChatGPT. They were asked to first attempt to use ChatGPT's Python interpreter module to conduct the analysis. However, they were allowed to run analysis code locally (in R or Stata) when ChatGPT failed to run the analysis itself. When running code locally, the teams were not allowed to use any other code except code provided by ChatGPT, with the exception that the teams could adjust file paths and their environment without the assistance of ChatGPT. We relied on the good nature of AI-led teams to *not* look at the studies, codes, or files. That is, we asked them to pass everything into ChatGPT. Last, we did not give specific guidance on how teams should operate. Teams could do the work independently or jointly throughout the event.

We have 103 teams; 33 human teams (92 researchers), 35 AI-assisted teams (93 researchers) and 35 AI-led teams (103 researchers). We show in Table S3 that the treatment arms are balanced across a large number of observables.

2.2 Three Tasks

We focus on three objective tasks. (*43*) First, teams were asked to computationally reproduce a few selected numerical results in the study assigned to them. Computational reproducibility involves using the same data as the original authors and running their codes. Teams had to fill out the time it took to computationally reproduce the numerical result. Of note, AB, JA and DM successfully computationally reproduced all results with minimal changes (e.g., changing paths) prior to the event.

Second, we asked teams to report any coding errors and data irregularities they found. One type of coding error would be discrepancies between the study and the code. We label the coding errors found as major or minor. We define coding errors as minor or major depending on whether the coding error could, in theory, have an impact on the claims tested. For instance, a coding error or data irregularity that impacts the dependent or independent variables could potentially have an impact on the estimation results. In contrast, a coding error for missing packages/paths or versioning issues is considered a minor error. Those coding errors are typically easily fixed by the reproducers and do not impact the validity of the claims made by the original authors. AB, JA and DM discussed all errors uncovered and classified coding errors as major or minor based on the general principles established above.

Third, we asked each team to report and perform two robustness checks. Qualifying and quantifying what makes a robustness check "good" or "bad" is not straightforward. We propose four different binary measures which we believe qualify a good robustness check: (i) clear (not vague) regarding purpose and execution; (ii) feasibility, (iii) not previously done by the original author(s); and (iv) focuses on the validity of the empirical strategy. Items (i) through (iii) are necessary conditions to be considered "good" robustness checks. Item (iv) we believe to be the purpose of robustness checks (especially): providing evidence which strengthens the credibility of the empirical strategy, from which conclusions of the study are being made. (*44–46*) A "good" robustness check therefore contains all of these elements. If one of these elements is missing, we classify it as a "bad" robustness check. In line with these considerations, we asked each team to propose two robustness checks that were not previously conducted by the authors and mentioned in the study or its supplementary materials. We further instructed participants that the robustness checks had to be feasible

and that heterogeneity analysis (e.g., comparing female and male respondents) was not considered a robustness check. AB, JA and DM classified robustness checks performed as being "good" or "bad".

3 Results

Our analyses were pre-registered after the pilot event in Toronto. We list deviations from our pre-analysis plan in the supplementary materials and note throughout whether the analysis is exploratory.

Computational Reproducibility

For computational reproducibility we have two different dependent variables: one as a binary (completed computational reproducibility versus did not complete reproducibility) and one which is continuous - time (in minutes) from the start of the event to when teams completed computational reproducibility. A completed computational reproduction is defined as having successfully ran the original authors' codes and produced numerical results identical to those in the article.

Our main finding is that computational reproducibility rates varied substantially across the groups. Most human (94%) and AI-assisted (91%) teams could computationally reproduce the results, while only 37% of AI-led teams could (Table 1). Table 2 shows our main regression estimates using OLS. See Table S4 for logit and poisson regressions and Table S5 for the control variables estimates. We find that human teams are about 59 percentage points more likely than AI-led teams to successfully computationally reproduce the results ($p < 0.01$). In contrast, there is no significant difference between human and AI-assisted teams ($p = 0.70$).

In Figure 1, we investigate whether the distribution of time to computational reproduction varies across groups. Note that many teams did not complete computational reproducibility leading to an important sample selection bias, especially for AI-led teams. We find that the median time for AI-led teams is 2 hours and 48 minutes. Most human teams, while faster than AI-led teams, take between one and three hours to complete the task (median 1 hour and 18 minutes). Most of the AI-assisted teams were done with the task in 1 hour and 45 minutes, with a few teams taking almost the entire event (median 1 hour and 12 minutes).

In an exploratory analysis, we investigate whether AI-led teams improved over time. In our setting, improvements would be due to the use of new ChatGPT versions or increased researchers' skills over time rather than learning between events. In Figure S1, we show the difference in computational reproducibility rates for each event. Visually, we observe that, in comparison to human teams, AI-led teams did not improve over time, with our first and last events both having a difference in reproducibility rate between human and AI-led teams of about 50 percentage points.

Coding Errors or Data Irregularities

We have two dependent variables, both of which are counts, of major and minor coding errors. We find that human teams identified on average 1.4 minor coding errors. In contrast, AI-assisted and AI-led teams respectively uncovered on average 0.94 and 0.5 minor coding errors (Table 1). We uncover a similar pattern for major errors, with human teams correctly identifying more errors. Table 2 confirms that the human teams uncovered significantly more errors than AI-assisted and AI-led teams. We also find that AI-assisted teams uncovered significantly more minor ($p = 0.022$) and major ($p = 0.018$) coding errors than AI-led teams. See Supplementary Materials Coding Errors and Data Irregularities for examples of coding errors and a discussion.

Figure 2 investigates whether the distribution of time to finding a first coding error varies across groups. The figure illustrates both the time to detecting a first minor and a first major error. Caution is required with this figure as human teams are more likely to detect coding errors than AI-assisted and AI-led teams. We provide weak evidence that AI-assisted teams are faster at uncovering a first coding error, and that AI-led teams are slower.

Our findings suggest that unaided human teams were more effective at detecting both major and minor errors compared to AI-led teams, highlighting a challenge in AI-led teams' ability to autonomously navigate and interpret complex code and data irregularities. We also find that human-only teams performed significantly better than AI-assisted teams on error detection, particularly in identifying errors with potentially significant implications (i.e., major errors). In an exploratory analysis, we show that this result is present only for teams working with Stata (Tables S6 and S7). The difference between human and AI-assisted teams could be due to many explanations including overreliance on AI – AI-assisted teams could be following ChatGPT's suggestions without seek-

ing and processing more information. (*36*) We also provide non-causal evidence in an exploratory analysis that AI-assisted teams with more AI experience uncovered more coding errors in Table S8, suggesting that AI training and practice may be critical for realizing the full benefits of AI assistance.

We investigate in an exploratory analysis if AI-led teams improved over time at detecting coding errors. In Figures S2 and S3, we provide weak evidence that AI-led teams are improving over time in detecting coding errors in comparison to human groups, possibly due to the use of new ChatGPT versions or increased researchers' skills over time.

Robustness Checks

We have four dependent variables. The first two are whether the reproducers proposed one or two "good" robustness checks. The third and fourth dependent variables are whether the reproducers implemented one or two "good" robustness checks.

We find that human and AI-assisted teams performed much better than AI-led teams. We find that all human and AI-assisted teams proposed at least one good robustness check, with 88% human teams suggesting two robustness checks in comparison to 86% for AI-assisted teams. In contrast, only 83% AI-led teams suggested one good and 63% suggested two good robustness checks. The difference for these two variables between AI-led teams and the other treatment arms are statistically significant (Tables 1 and 2). We find similar significant differences for the implementation of "good" robustness checks, with AI-led teams being about 33–37 percentage points less likely to implement "good" robustness checks. The main criterion leading AI-led teams to suggest "bad" robustness checks is that they were already implemented by the original author(s). Further, six AI-led teams did not provide any robustness check.

Our results indicate that AI-led teams, while able to produce checks with some level of quality, faced more challenges in aligning with the criteria, potentially due to limited human guidance in interpreting the empirical strategy and ensuring feasibility.

Overreliance or Underreliance of AI for AI-Assisted

We investigate whether there is a relationship between AI use and performance for AI-assisted teams in Table S9. Teams were divided into two groups based on the median total number of prompts used. This analysis is exploratory and should not be viewed as causal. We find that AI-assisted teams that used less AI were less likely to computationally reproduce the results. However, they identified more minor and major coding errors and took less time to computationally reproduce and find their first (minor and major) coding error. These results suggest that some AI-assisted teams may have overrelied on AI support. (*40, 41, 47–49*)

Discussion

The findings of this study offer critical insights into the potential and limitations of AI-assisted and AI-driven approaches in the reproducibility of empirical social science research. Computational reproducibility, error detection, and robustness checks are essential components of empirical research validation, and assessing these through the lenses of human-only, AI-assisted, and AI-led teams sheds light on how AI may be integrated in the expensive reproduction process, accelerating it and improving its overall reliability. Although recent advancements in LLMs have opened possibilities for AI integration in research (*50, 51*), our results suggest that, while AI-driven reproduction has potential to save time and money for a subset of studies, a significant human component remains crucial in ensuring successful computational reproduction for most studies. The optimal role for AI in reproducibility may therefore still be as a collaborator for most studies rather than a sole executor. LLM systems could be used as a first pass helping to identify coding errors (*52*) and proposing possible solutions to them (*53, 54*), while on a subsequent step humans would still play the pivotal role of performing a more in-depth evaluation.

Summary of Findings

AI-led teams have faced notable challenges compared to both AI-assisted and human-only teams. Only 37% of AI-led teams were able to successfully complete computational reproducibility, highlighting a substantial gap in the current capacity of AI to autonomously navigate complex quantitative analyses. Similarly, in error detection, AI-led teams detected significantly fewer major and

minor errors than either AI-assisted or human teams. These errors, particularly major ones with implications for research claims, often require nuanced understanding and critical thinking to identify, a capability that AI tools in their current form lack (see Supplementary Materials Coding Errors and Data Irregularities). These results suggest that AI's role may be best suited for supporting tasks where direct interpretative judgment is less critical, or where error detection can be supplemented by human oversight.

Limitations

One limitation is our focus on solely OpenAI's ChatGPT using GPT-4/4o models, meaning that we cannot generalize to all current AI models. Furthermore, the limited timeframe of seven hours for study teams to complete their reproductions may not adequately reflect the conditions under which reproducibility efforts are typically conducted. Finally, we relied on a small number of research papers illustrating a relatively narrow range of social science methodologies and techniques, which makes it difficult to generalize our findings to work with AI systems across all social science subfields.

Implications for Human-AI Collaboration in Research

Our findings support the notion that, while AI tools hold promise for aiding in reproducibility tasks, the state of technology as of 2024 is not yet advanced enough for full autonomy in complex empirical workflows. Human expertise remains critical to navigate challenges and provide interpretative guidance for reproducibility and error detection. The AI-assisted model—where humans work alongside AI tools— did not emerge as a winner over humans only teams.

In scenarios where computational reproducibility, error detection, and robustness checks require in-depth understanding, domain knowledge, and flexible problem-solving, human involvement currently adds value. The ability to contextualize, interpret, and implement complex quantitative research remains a human strength, underscoring the current limitations of AI in fully autonomous reproduction efforts.

Future Outlook

Looking ahead, future advancements in models optimized through reinforcement learning to solve reasoning problems using a chain of thought could address the limitations we reported, possibly improving the model's ability to reproduce complex quantitative research through iterative, reasoning-driven processes. As LLMs evolve to incorporate better contextual understanding and reasoning, their role in reproducibility tasks could shift from support to a more central position, especially in less complex, structured reproduction settings. Future iterations of AI tools may incorporate improvements in interpreting code and data irregularities, detecting nuanced errors, and generating plausible robustness checks with minimal human input. Such advancements could enhance AI's ability to autonomously execute reproducibility tasks, reducing the reliance on human oversight for routine or straightforward reproducibility challenges.

Future research should consider the potential for training models specifically on social science and quantitative research contexts. Current LLMs are trained on vast datasets but may lack specificity in understanding the unique demands of empirical social science research. AI systems tailored for social science reproduction could potentially improve reproducibility outcomes, reducing the barriers AI currently faces in autonomously handling the nuances of quantitative research. Additionally, incorporating continuous feedback and learning mechanisms could allow AI-assisted and AI-led teams to improve performance over time, as AI learns from each reproduction task and adapts based on human feedback.

Table 1: Comparison of Human, AI-Assisted, and AI-Led Metrics

Variable	Human-Only	AI-Assisted	AI-Led	Human-Only vs AI-Assisted	Human-Only vs AI-Led	AI-Assisted vs AI-Led
Reproduction	0.939 (0.242)	0.914 (0.284)	0.371 (0.490)	0.025 (0.697)	0.568 (<0.001)	0.543 (<0.001)
Minutes to reproduction	79.1 (43.6)	92.3 (88.3)	175.1 (65.6)	-13.3 (0.454)	-96.0 (<0.001)	-82.7 (0.004)
Number of minor errors	1.424 (1.696)	0.943 (1.454)	0.514 (0.919)	0.481 (0.213)	0.910 (0.007)	0.429 (0.145)
Minutes to first minor error	95.7 (77.5)	74.0 (46.5)	150.1 (107.5)	21.7 (0.318)	-54.4 (0.109)	-76.1 (0.016)
Number of major errors	1.364 (1.496)	0.629 (0.942)	0.229 (0.490)	0.735 (0.017)	1.135 (<0.001)	0.400 (0.029)
Minutes to first major error	147.2 (89.4)	128.8 (53.4)	191.7 (97.6)	18.4 (0.496)	-44.5 (0.278)	-62.9 (0.069)
At least one good robustness check	1.000 (0.000)	1.000 (0.000)	0.829 (0.382)	- (-)	0.171 (0.012)	0.171 (0.010)
At least two good robustness checks	0.879 (0.331)	0.857 (0.355)	0.629 (0.490)	0.022 (0.796)	0.250 (0.017)	0.229 (0.029)
Ran at least one good robustness check	0.939 (0.242)	0.943 (0.236)	0.571 (0.502)	-0.003 (0.953)	0.368 (<0.001)	0.371 (<0.001)
Ran At least two good robustness checks	0.788 (0.415)	0.800 (0.406)	0.457 (0.505)	-0.012 (0.903)	0.331 (0.005)	0.343 (0.003)

Note: Standard errors in parentheses for individual branches (Human-only, AI-Assisted, and AI-Led); p-values in parentheses for branch comparisons (Human-Only Vs AI-Assisted, Human-Only Vs AI-Led, and AI-Assisted Vs AI-Led).

Table 2: Results from OLS regressions predicting reproduction outcomes

	(1) Reproduction	(2) Minor errors	(3) Major errors	(4) One good robustness	(5) Two good robustness	(6) Ran one robustness	(7) Ran two robustness
AI-Assisted	-0.018	-0.487*	-0.646**	-0.009	-0.014	-0.032	-0.009
	(0.063)	(0.270)	(0.254)	(0.027)	(0.103)	(0.061)	(0.113)
	[-0.144; 0.107]	[-1.025; 0.051]	[-1.153; -0.139]	[-0.063; 0.046]	[-0.220; 0.191]	[-0.155; 0.090]	[-0.233; 0.216]
AI-Led	-0.593***	-1.050***	-1.136***	-0.167**	-0.250**	-0.323***	-0.290**
	(0.090)	(0.258)	(0.235)	(0.068)	(0.107)	(0.098)	(0.126)
	[-0.773; -0.413]	[-1.565; -0.536]	[-1.604; -0.667]	[-0.302; -0.031]	[-0.463; -0.037]	[-0.518; -0.127]	[-0.540; -0.040]
Controls	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.738	0.951	0.728	0.942	0.786	0.816	0.680
p-val (AI-Assisted vs. AI-Led)	0.000	0.022	0.018	0.023	0.036	0.004	0.019
Observations	103	103	103	103	103	103	103

Note: Standard errors in parentheses, confidence intervals in brackets; human-only branch omitted.

Controls include number of teammates; game-software, skill, and attendance fixed effects.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

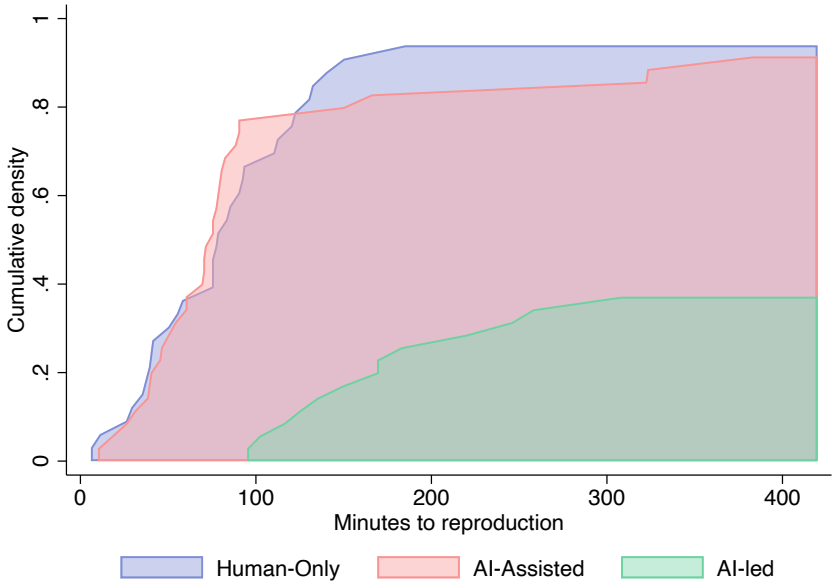

Figure 1: Cumulative density plot depicting the time (in minutes) taken by each treatment arm—Human, AI-assisted, and AI-Led—to computationally reproduce the main findings in the assigned study.

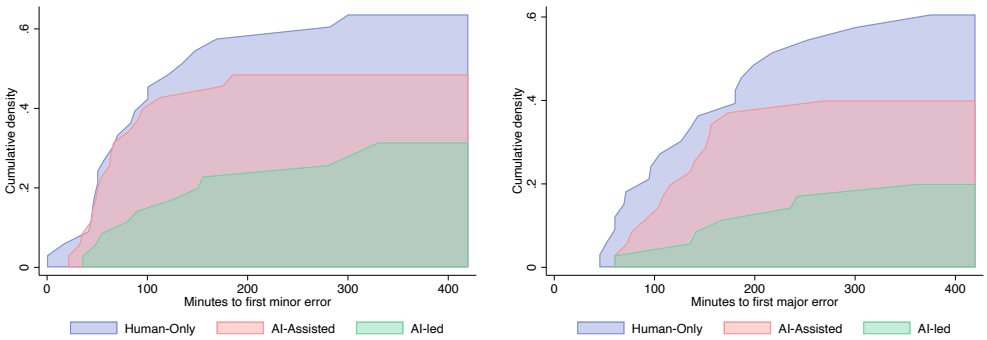

Figure 2: Cumulative density plots showing the time (in minutes) taken by each treatment arm—Human, AI-assisted, and AI-Led—to identify the first minor (left) and major (right) coding errors in the assigned study.

References and Notes

1. A. Brodeur, *et al.*, Promoting reproducibility and replicability in political science. *Research & Politics* **11** (1), 20531680241233439 (2024).

2. D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, V. Stodden, Reproducible research in computational harmonic analysis. *Computing in Science & Engineering* **11** (1), 8–18 (2008).

3. M. Fišar, *et al.*, Reproducibility in Management Science. *Management Science* **70** (3), 1343–1356 (2024).

4. P. Gertler, S. Galiani, M. Romero, How to Make Replication the Norm. *Nature* **554** (7693), 417–9 (2018).

5. S. N. Goodman, D. Fanelli, J. P. Ioannidis, What does research reproducibility mean? *Science Translational Medicine* **8** (341), 341ps12–341ps12 (2016).

6. A. Marcoci, *et al.*, Predicting the replicability of social and behavioural science claims from the COVID-19 Preprint Replication Project with structured expert and novice groups, metaArXiv Preprint.

7. M. Milkowski, W. M. Hensel, M. Hohol, Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience* **45** (3), 163–172 (2018).

8. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (National Academies Press) (2019), doi:10.17226/25303, `https://www.nap.edu/catalog/25303`.

9. C. Pérignon, K. Gadouche, C. Hurlin, R. Silberman, E. Debonnel, Certify reproducibility with confidential data. *Science* **365** (6449), 127–128 (2019).

10. L. Vilhuber, Reproducibility and replicability in economics. *Harvard Data Science Review* **2** (4), 1–39 (2020).

11. L. Vilhuber, Report by the AEA Data Editor. *AEA Papers and Proceedings* **112**, 813–23 (2022), doi:10.1257/pandp.112.813.

12. A. Brodeur, *et al.*, Mass Reproducibility and Replicability: A New Hope (2024), Institute for Replication Discussion Paper 107.

13. A. C. Chang, P. Li, Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not". *Critical Finance Review* **11** (1), 185–206 (2022).

14. S. Crüwell, *et al.*, What's in a badge? A computational reproducibility investigation of the open data badge policy in one issue of Psychological Science. *Psychological Science* **34** (4), 512–522 (2023).

15. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349** (6251), aac4716 (2015).

16. P. Obels, D. Lakens, N. A. Coles, J. Gottfried, S. A. Green, Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science* **3** (2), 229–237 (2020).

17. C. Pérignon, *et al.*, Computational reproducibility in finance: Evidence from 1,000 tests. *The Review of Financial Studies* **37** (11), 3558–3593 (2024).

18. V. Stodden, J. Seiler, Z. Ma, An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences* **115** (11), 2584–2589 (2018).

19. B. Wood, R. Müller, A. Brown, Push Button Replication: Is Impact Evaluation Evidence for International Development Verifiable? (2018), `https://osf.io/n7a4d/`, OSF Preprints.

20. W. X. Zhao, *et al.*, A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).

21. Y. Chang, *et al.*, A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* **15** (3), 1–45 (2024).

22. G. Bansal, *et al.*, Does the whole exceed its parts? The effect of AI explanations on complementary team performance, in *Proceedings of the 2021 CHI conference on human factors in computing systems* (2021), pp. 1–16.

23. G. Bansal, B. Nushi, E. Kamar, E. Horvitz, D. S. Weld, Is the most accurate ai the best teammate? Optimizing AI for teamwork, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35 (2021), pp. 11405–11414.

24. E. Bondi, *et al.*, Role of human-AI interaction in selective prediction, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36 (2022), pp. 5286–5294.

25. Á. A. Cabrera, A. Perer, J. I. Hong, Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction* **7** (CSCW1), 1–21 (2023).

26. E. Goh, *et al.*, Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. *medRxiv* (2024).

27. B. Koepnick, *et al.*, De novo protein design by citizen scientists. *Nature* **570** (7761), 390–394 (2019).

28. H. Liu, V. Lai, C. Tan, Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* **5** (CSCW2), 1–45 (2021).

29. H. Mozannar, *et al.*, Effective human-AI teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems* **36** (2024).

30. S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381** (6654), 187–192 (2023).

31. C. Reverberi, *et al.*, Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports* **12** (1), 14952 (2022).

32. M. Schemmer, P. Hemmer, M. Nitsche, N. Kühl, M. Vössing, A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making, in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), pp. 617–626.

33. M. H. Tessler, *et al.*, AI can help humans find common ground in democratic deliberation. *Science* **386** (6719), eadq2852 (2024).

34. A. Toner-Rodgers, Artificial Intelligence, Scientific Discovery, and Product Innovation (2024), preprint on Author's webpage.

35. M. Vaccaro, J. Waldo, The effects of mixing machine learning and human judgment. *Communications of the ACM* **62** (11), 104–110 (2019).

36. M. Vaccaro, A. Almaatouq, T. Malone, When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* pp. 1–11 (2024).

37. B. Wilder, E. Horvitz, E. Kamar, Learning to complement humans, in *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (2020), pp. 1526–1533.

38. Cherry Bekaert LLP, Report of Independent Auditor (2022), doi:10.1257/aer.112.6.2083, `https://pubs.aeaweb.org/doi/10.1257/aer.112.6.2083`.

39. J.-E. Colliard, C. Hurlin, C. Perignon, The Economics of Computational Reproducibility (2022), SSRN: https://ssrn.com/abstract=3418896.

40. Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* **5** (CSCW1), 1–21 (2021).

41. L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making? *International Journal of Human-Computer Studies* **51** (5), 991–1006 (1999).

42. OpenAI, *et al.*, GPT-4 Technical Report (2024), `https://arxiv.org/abs/2303.08774`.

43. Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in *Proceedings of the 25th International Conference on Intelligent User Interfaces* (2020), pp. 454–464.

44. M. Del Giudice, S. W. Gangestad, A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science* **4** (1), 2515245920954925 (2021).

45. X. Lu, H. White, Robustness checks and robustness tests in applied economics. *Journal of Econometrics* **178**, 194–206 (2014).

46. M. B. Nuijten, Assessing and improving robustness of psychological research findings in four steps, in *Avoiding questionable research practices in applied psychology* (Springer), pp. 379–400 (2022).

47. V. Lai, H. Liu, C. Tan, "Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13.

48. Y. Zhang, Q. V. Liao, R. K. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), pp. 295–305.

49. H. Vasconcelos, *et al.*, Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* **7** (CSCW1), 1–38 (2023).

50. Y. K. Dwivedi, *et al.*, Opinion Paper:"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* **71**, 102642 (2023).

51. B. D. Lund, *et al.*, ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* **74** (5), 570–581 (2023).

52. N. Wadhwa, *et al.*, Core: Resolving code quality issues using llms. *Proceedings of the ACM on Software Engineering* **1** (FSE), 789–811 (2024).

53. Y. Zhang, Detecting code comment inconsistencies using llm and program analysis, in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering* (2024), pp. 683–685.

54. D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, B. Myers, Using an llm to help with code understanding, in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (2024), pp. 1–13.

55. Open AI, File uploads FAQ (2024), https://help.openai.com/en/articles/8555545-file-uploads-faq/ [Accessed: November 28, 2024].

56. Open AI, Learning to Reason with LLMs (2024), https://openai.com/index/learning-to-reason-with-llms/ [Accessed: November 18, 2024].

Acknowledgments

We would like to thank Gabriel Zimmerman for research assistance.

Funding: This research and AI replications games were funded by Open Philanthropy project "Benchmarking LLM agents on real-world tasks: Reproducibility" and the Alfred P. Sloan Foundation Foundation grant G-2023-22326. We also benefited from funding to host games from the Universities of Toronto, Ottawa, Cornell and Tilburg. Mahmoud Elsherif acknowledges funding from Leverhulme Early Career Research Fellowship-ECF-2022-761. Shumi Akhtar acknowledges funding DP200102935 awarded by the Australian Research Council Grant.

Author contributions: AB, DV, AM, JPA, DM, BB, RA: conception of the work. AB, GB, DV, JPA: analysis and interpretation of data. AB, DV, AM, JPA, DM, BB wrote the original draft, lead revision, and take responsibility for the content – while SA, MA, CB, NC, GB, MB, FC, LD, IEF, LF, JF, JG, GG, FH, MH, ZI, DL, BM, RM, XM, AO, PP, NR, OS, GT, RT, RS, TS, AV and LW contributed to editing and review through commenting on drafts of the paper. DV: AI training. All coauthors except AB, DV, AM, DM, BB, RA, TS: participation in replication games.

Competing interests: The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada. AM is a UKRI Policy Fellow seconded to the Department for Science, Innovation and Technology. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department for Science, Innovation and Technology or the UK Government.

Data and materials availability: We make our (i) AI training materials and recording, (ii) data and codes, (iii) pre-analysis plan and (iv) template form available here: `https://osf.io/sz2g8/`. We declare no restrictions on sharing or re-use.

Supplementary materials

Materials and Methods

Figs. S1 to S3

Tables S1 to S8

Supplementary Materials for Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science

Abel Brodeur (University of Ottawa; Institute for Replication), David Valenta (University of Ottawa), Alexandru Marcoci (University of Nottingham, University of Cambridge), Juan P. Aparicio (University of Ottawa; Institute for Replication), Derek Mikola (University of Ottawa; Institute for Replication), Bruno Barbarioli (University of Ottawa; Institute for Replication), Rohan Alexander (University of Toronto), Lachlan Deer (Tilburg University), Tom Stafford (Sheffield University), Lars Vilhuber (Cornell University), Gunther Bensch (RWI - Leibniz Institute for Economic Research), Mohamed Abdelhady (Carleton University), Yousra Abdelmoula (Carleton University), Ghina Abdul Baki (University of Ottawa), Tomás Aguirre (Centre for the Governance of AI), Sriraj Aiyer (University of Oxford), Shumi Akhtar (The University of Sydney), Farida Akhtar (Macquarie University), Melle R. Albada (Vienna University of Economics and Business), Micah Altman (MIT), David Angenendt (Technical University of Munich), Zahra Arjmandi Lari (Independent researcher), Jorge Armando De León Tejada (Universidad del Rosario), Igor Asanov (International Center for Higher Education Research and Faculty of Economics, University of Kassel), Anastasiya-Mariya Asanov Noha (University of Kassel, INCHER), Rebecca Ashong (University of Ghana), Tobias Auer (London School of Economics), Francisco J. Bahamonde-Birke (Tilburg University), Bradley J. Baker (Temple University), Söhnke M. Bartram (University of Warwick and CEPR), Dongqi Bao (University of Zurich), Lucija Batinovic (Linköping University), Tommaso Batistoni (University of Oxford), Monica Beeder (NHH Norwegian School of Economics), Louis-Philippe Beland (Carleton University), Carsten Bienz (Norwegian School of Economics), Christ Billy Aryanto (Faculty of Psychology, Atma Jaya Catholic University of Indonesia), Cylcia Bolibaugh (University of York), Carl Bonander (University of Gothenburg), Ramiro Bravo (Faculty of Biology, Medicine and Health, The University of Manchester), Katherine Brennan (Bank of Canada), Egor Bronnikov (Maastricht University; George Mason University), Stephan Bruns (Hasselt University), Nino Buliskeria (Nazarbayev University), Sara Caicedo-Silva (Universidad de

los Andes), Andrea Calef (University College London, School of Management), Solomon Caulker (United Methodist University Sierra Leone), Simonas Cepenas (ISM University of Management and Economics), Arthur Chatton (Université Laval), Zirou Chen (University of Toronto), Ngozi Chioma Ewurum (Michael Okpara University of Agriculture, Umudike, Nigeria), Anda-Bianca Ciocîrlan (University of Sheffield), Felix J. Clouth (Tilburg University), Jason Collins (University of Technology Sydney), Nikolai Cook (Wilfrid Laurier University), Cesar Cornejo (The London School of Hygiene & Tropical Medicine), João Craveiro (University of Sheffield), Jing Cui (University of Ottawa), Niveditha Chalil Vayalabron (School of Earth and Planetary Science, National Institute of Science Education and Research, India), Christian Czymara (Goethe-Universitaet Frankfurt), Carlos Daniel Bermúdez Jaramillo (Universidad del Rosario), Hannes Datta (Tilburg University), Lien Denoo (Tilburg University), Arshia Dhaliwal (Carleton University), Nency Dhameja (Binghamton University), Elodie Djemai (Université Paris-Dauphine), Erwan Dujeancourt (Stockholm University), Uğurcan Dündar (Vienna University of Economics and Business), Thibaut Duprey (Bank of Canada), Yasmine Eissa (The American University in Cairo), Youssef El Fassi (HEC Lausanne), Ismail El Fassi (University of St. Gallen), Keaton Ellis (UC Berkeley), Ali Elminejad (Nazarbayev University), Mahmoud Elsherif (University of Leicester), Aysil Emirmahmutoglu (NHH Norwegian School of Economics), Giulian Etingin-Frati (University of Zurich), Emeka Eze (Michael Okpara University of Agriculture), Jan Fabian Dollbaum (University College Dublin), Jan Feld Victoria (University of Wellington), Andres Felipe Rengifo Jaramillo (Business School; Universidad de los Andes), Guidon Fenig (University of Ottawa), Victoria Fernandes (Bank of Canada), Lenka Fiala (University of Bergen), Lukas Fink (FU Berlin), Sara Fish (Harvard University), Jack Fitzgerald (Vrije Universiteit Amsterdam), Rachel Forshaw (Heriot-Watt University), Alexandre Fortier-Chouinard (Université Laval), Louis Fréget (CEPREMAP), Joris Frese (European University Institute), Jacopo Gabani (World Bank; Centre for Health Economics, University of York), Sebastian Gallegos (UAI Business School), Max C. Gamill (University of Sheffield), Attila Gáspár (HUN-REN Centre for Economic and Regional Studies), Romain Gauriot (Deakin University), Evelina Gavrilova (NHH Norwegian School of Economics), Diogo Geraldes (University College Dublin), Giulio Giacomo Cantone (University of Sussex), Grant Gibson (McMaster University & CRDCN), Dirk Goldschmitt (University of Sheffield), Amélie Gourdon-Kanhukamwe (King's College London), Andrea Gregor de Varda (University of Milano-Bicocca), Idaliya Grigo-

ryeva (UC San Diego), Alexi Gugushvili (University of Oslo), Aaron H.A. Fletcher (University of Sheffield), Florian Habermann (University of Lausanne), Márton Hablicsek (Leiden University), Joanne Haddad (Université Libre de Bruxelles), Jonathan D. Hall (University of Alabama), Olle Hammar (Linnaeus University and Institute for Futures Studies), Malek Hassouneh (University of Toronto), Carina I Hausladen (ETH Zürich), Sophie C. F. Hendrikse (Tilburg University), Matthew Hepplewhite (University of Oxford), Anson T. Y. Ho (Toronto Metropolitan University), Senan Hogan-Hennessy (Cornell University), Elliot Howley (University of Nottingham), Gaoyang Huang (Swiss Federal Institute of Technology in Zurich), Héloïse Hulstaert (Hasselt University; Liège University), Zlatomira G. Ilchovska (University of York; University of Birmingham), Niklas Jakobsson (Karlstad University), Joakim Jansson (Linnaeus University; Research Institute of Industrial Economics), Ewa Jarosz (University of Warsaw), Hossein Jebeli (Bank of Canada), Yanchen Jiang (Harvard University), Hiba Junaid (Bart's Life Sciences, Bart's Health NHS Trust; Queen Mary university of London), Rohan Kalluraya (Cornell University), Edmund Kelly (University of Oxford), Eva Kimel (University of York), Sorravich Kingsuwankul (Vrije Universiteit Amsterdam), Valentin Klotzbücher (University of Freiburg), Daniel Krähmer (University of Munich), Pijus Krūminas (ISM University of Management and Economics), Nicholas Kruus (Schelling Research), Essi Kujansuu (University of Innsbruck), Christoph F. Kurz (Ludwig-Maximilians-Universität Munich) Stephan Küster (Freie Universität Berlin), Blake Lee-Whiting (University of Toronto), Felix Lewandowski (University of Nottingham), Tongzhe Li (University of Guelph), Ruoxi Li (Yale University), Dan Liu (Australian National University), Jiacheng Liu (Purdue University), Helix Lo (University of Tokyo), Katharina Loter (Tilburg University), Felipe Macedo Dias (Cornell University), Christopher R. Madan (University of Nottingham), Nicolas Mäder (University of San Diego), Marco Mandas (University of Cagliari), Jan Marcus (FU Berlin), Diego Marino Fages (Durham University), Xavier Martin (Tilburg University), Ryan McWay (University of Minnesota), Daniel Medina-Gaspar (Universidad EAFIT), Sisi Meng (University of Notre Dame), Lingyu Meng (University of Sheffield), Alex P. Miller (University of Southern California), Thibault Mirabel (Equalis Capital), Dibya Deepta Mishra (Rice University), Sumit Mishra (Krea University), Belay W. Moges (Dilla University), Morteza Mohandes Mojarrad (Tilburg university), Myra Mohnen (University of Ottawa), Louis-Philippe Morin (University of Ottawa), Fabio Motoki (University of East Anglia), Lucija Muehlenbachs (University of Calgary), Gastón Mullin (Tilburg University), Andreea

Musulan (University of Montreal), Sara Muzzì (University of Milano Bicocca), James A. C. Myers (University of Sheffield), Florian Neubauer (RWI - Leibniz Institute for Economic Research), Tuan Nguyen (Hasselt University), Ali Niazi (University of Calgary), Ardyn Nordstrom (Carleton University), Bartłomiej Nowak (Cardinal Stefan Wyszynski University), Daneal O'Habib (Bank of Canada), Tim Ölkers (University of Göttingen), Justin Ong (University of Sheffield), Valeria Orozco Castiblanco (IESE, Universidad de Navarra), Ömer Özak (SMU), Ali I. Ozkes (SKEMA Business School, GREDEG, Université Côte d'Azur), Mikael Paaso (Erasmus University Rotterdam), Shubham Pandey (Universität Osnabrück), Varvara Papazoglou (University of Sheffield), Romeo Penheiro (University of Houston), Linh Pham (Lake Forest College), Ulrike Phieler (Vienna University of Economics and Business), Peter Pütz (Bielefeld University), Quan Qi (University at Albany, SUNY), Jingyi Qiu (University of Michigan), David A. Reinstein (The Unjournal), Juuso Repo (INVEST Flagship Research Center, University of Turku), Nicolas Rudolf (University of Lausanne), Shree Saha (Cornell University), Orkun Saka (City, University of London), Chiara Saponaro (University of Milano-Bicocca), Georg Sator (University of Nottingham), Martijn Schoenmakers (Tilburg University), Raffaello Seri (InsIDE Lab, DiEco, Università degli Studi dell'Insubria), Meet Shah (Toronto Metropolitan University), Paul Sibille (University of Liege), Christoph Siemroth (University of Essex), Vladimir Skavysh (Bank of Canada), Ben Slater (University of Cambridge), Wenting Song (Bank of Canada), Stefan Staubli (University of Calgary), Tobias Steindl (University of Regensburg), Nomwendé Steven Waongo (University of Ottawa), Paul Stott (University of Manchester), Stephenson Strobel (McMaster University), Roshini Sudhaharan (Tilburg University), Pu Sun (University of Ottawa), Scott D. Swain (Clemson University), Oleksandr Talavera (University of Birmingham), Hanz M. Tantiangco (University of Sheffield), Georgy Tarasenko (Cornell University), Boyd Tarlinton (Department of Primary Industries, QLD), Mariam Tarraf (Carleton University), Ken Teoh (International Monetary Fund), Rémi Thériault (Université du Québec à Montréal), Bethan Thompson (SRUC), Tonghui Tian (Carleton University), Wenjie Tian (University of Ottawa), Manuel Tobias Rein (Tilburg University), Emmanuel Tolani (University of Bonn), Nicolai Topstad Borgen (University of Oslo), Solveig Topstad Borgen (University of Oslo), Javier Torralba (Tilburg University), Carolina Velez-Ospina (World Bank), Man Wai Mak (Carleton University), Lukas Wallrich (Birkbeck, University of London), Zeyang Wang (Vanderbilt University), Leah Ward (University of Manchester), Matthew D. Webb (Carleton University), Duncan Webb

(Princeton University), Bryan S. Weber (College of Staten Island, CUNY), Christoph Weber (ESSCA School of Management), Wei-Chien Weng (National Taiwan University), Christian Westheide (University of Vienna), Tom Wilkinson (University of Sheffield), Kwong-Yu Wong (National University of Singapore), Marcin Wroński (Collegium of World Economy, SGH Warsaw School of Economics), Zhuangchen Wu (University of Birmingham), Qixia Wu (University of Ottawa), Victor Y. Wu (Stanford University), Bohan Xiao (University of Ottawa), Feihong Xu (Northwestern University), Cong Xu (National Chengchi University; Aalto University), Pranav Yadav (Tilburg University), Yu Yang Chou (University College London), Luther Yap (Princeton University), Myra Yazbeck (University of Ottawa), Zuzanna Zagrodzka (University of Sheffield), Tahreen Zahra (Carleton University), Mirela Zaneva (University of Oxford), Xiaomeng Zhang (Nanjing Audit University), Ziwei Zhao (University of Lausanne; Swiss Finance Institute), Han Zhong (University of Toronto), Aras Zirgulis (ISM University of Management and Economics), Jiacheng Zou (Columbia University), Floris Zoutman (NHH Norwegian School of Economics), Christelle Zozoungbo (Penn State University).

This PDF file includes:

Materials and Methods

Figures S1 to S3

Tables S1 to S8

Materials and Methods

Pre-Registration

Our pre-analysis plan was pre-registered on OSF on May 2nd, 2024: `https://osf.io/sz2g8/`. The pre-registration was done after our pilot event at the University of Toronto.

Of note, the pre-analysis plan refers to AI-assisted teams as cyborg teams and AI led teams as machine teams.

Research Questions

Here are the primary research questions that were pre-registered:

1. Do AI-led teams computationally reproduce more results than AI-assisted and human teams?
2. Are AI-led teams faster to computationally reproduce results than AI-assisted and human teams?
3. Do AI-led teams detect more major and minor coding errors or data irregularities than AI-assisted and human teams?
4. Are AI-led teams faster at detecting major and minor coding errors or data irregularities than AI-assisted and human teams?
5. Do AI-led teams detect more major and minor coding errors or data irregularities than AI-assisted and human teams?
6. Do AI-led teams propose better robustness checks than AI-assisted and human teams?
7. Are AI-led teams more capable of implementing robustness checks than AI-assisted and human teams?
8. Do AI-assisted teams computationally reproduce more results than human teams?
9. Are AI-assisted teams faster to computationally reproduce results than human teams?
10. Do AI-assisted teams detect more major and minor coding errors or data irregularities than human teams?
11. Are AI-assisted teams faster at detecting major and minor coding errors or data irregularities than human teams?
12. Do AI-assisted teams detect more major and minor coding errors or data irregularities than human teams?
13. Do AI-assisted teams propose better robustness checks than human teams? 14. Are AI-assisted

teams more capable of implementing robustness checks than human teams?

We also explored the following exploratory (pre-registered) research questions:

15. Are AI-led teams improving their performance over time at computationally reproducing results, detecting coding errors or data irregularities, and providing good robustness checks?"

16. Are AI-assisted teams improving their performance over time at computationally reproducing results, detecting coding errors or data irregularities, and providing good robustness checks?

We also tackle an exploratory research question that was not pre-registered in the article:

17. Do AI-assisted teams overrely or underrely on AI?

AI Replication Games Advertisement

The Institute for Replication advertised the AI replication games through social media (Bluesky and X) and emails. Events were also promoted on the Institute's webpage (`https://i4replication.org/games.html`) Only graduate students, postdoctoral fellows, professors and researchers from non-academic organizations with a PhD could register. All participants were promised coauthorship to this paper.

The typical social media posts included the following information:

"This is a one-day event that brings researchers together to collaborate on reproducing quantitative results published in high-ranking social science journals. You will have the opportunity to network with fellow researchers and develop your coding and AI skills.

Open to all researchers: faculty, post-docs, and graduate students. Knowledge of Python or R or Stata is essential. Participants will be randomly assigned to one of three teams: Machine with restricted human assistance, Cyborg or Human.

All participants will get coauthorship on a meta-research journal paper which combines the work of all teams.

Register here: Link to Registration Form."

Participants Exclusion

We did not accept registration from participants with no knowledge of Stata nor R. We also excluded from participating a very small number of researchers with no knowledge of R and who did not have a Stata license.

As noted in the main text, a few organizers participated in one of the games. They did not know about the papers to be reproduced at their respective event.

Documents to be Filled During the Games

During the event, each team filled out an excel sheet documenting their outcomes. See the excel document here: `https://osf.io/sz2g8/`. The document "Template Time Stamp" includes 3 sheets to be filled by each team. The first sheet is for computational reproducibility. Teams need to fill out the time that they have computationally reproduced the exhibit. The second sheet documents coding errors detected. Teams need to add a row for each coding error and data irregularity and enter the time they have detected them. In the last sheet, teams need to provide a description of their two robustness checks and provide estimates if they could implement those. Researchers in the AI-assisted and AI-led groups are also asked to share their prompts/conversations at the end of the event.

Study Selection

For each event, two studies published in leading social science journals are selected by AB. The studies are published in a journal with a data and code availability policy. One study is coded in Stata; the other is coded in R. The studies have all been reproduced by the Institute for Replication before the AI replication games. The Institute for Replication runs about two "regular" replication games, in contrast to these events, each month. At every such event, teams of researchers try to reproduce results from peer-reviewed publications. They then prepare reports of their findings which are subsequently shared with the original authors and made public on average six months following an event. Importantly, this means the Institute for Replication had over 20 published studies with known reproduction results *but have not yet been made publicly available* to choose from at any point in 2024. We could not take studies with publicly known coding issues since ChatGPT may be

able to "know" coding errors or data discrepancies without "finding" them. This is the corpus we sampled from for each of the AI replication games.

The sampling cannot be random or blind for a few reasons. First, the variation in reproduction packages (sometimes called *replication* packages in the social sciences) is too large. In both scenarios, where folders reproduce studies perfectly or folders that cannot be deciphered at all, would yield no variation in at least one of our outcomes. (No coding errors exist in the former; we can't evaluate the correctness of the code in the latter.) Second, studies need to rely on publicly available data and codes or the exercise is futile. Third, we need to match the software abilities of participants to each study. Within this corpus, we selected studies known to have coding errors or data irregularities. All teams were told that they needed to uncover coding errors or data irregularities.

Some studies were used for two events. The following studies were selected:

Pilot: Toronto Replication Games (with virtual researchers in Europe):

1-X. Labandeira et al., "Major Reforms in Electricity Pricing: Evidence from a Quasi-Experiment", The Economic Journal, (2022), vol.132(May): 1517–1541, DOI: 10.1093/ej/ueab076.

2-P. Christensen and C. Timmins, "Sorting or Steering: The Effects of Housing Discrimination on Neighborhood Choice", Journal of Political Economy, (2022), vol.130(August): 2110–2163, DOI: 10.1086/720140.

Materials: Documents shared on Dropbox with participants. We did not provide the screenshots for this pilot event.

Ottawa Replication Games:

1-X. Labandeira et al., "Major Reforms in Electricity Pricing: Evidence from a Quasi-Experiment", The Economic Journal, (2022), vol.132(May): 1517–1541, DOI: 10.1093/ej/ueab076.

2-Wolfowicz et al., "Arrests and Convictions but Not Sentence Length Deter Terrorism in 28 European Union Member States" Nature Human Behaviour, vol.7: (2023), 1878–1889, DOI: 10.1038/s41562-023-01695-6.

Materials: `https://osf.io/5v2km/`

Sheffield Replication Games:

1-P. Atanasov et al., "Taste-Based Gender Favouritism In High-Stake Decisions: Evidence from the Price is Right", The Economic Journal, (2023), vol.134(February): 856-883, DOI: 10.1093/ej/uead087.

2-R. Bajo-Buenestado and M. A. Borrella-Mas, "The Heterogeneous Tax Pass-Through Under

Different Vertical Relationships", The Economic Journal, (2022), vol.132(July): 1684–1708, DOI: 10.1093/ej/ueac007.

Materials: `https://osf.io/z48ax/`

Cornell Replication Games:

1-S. Hill and M. E. Roberts, "Acquiescence Bias Inflates Estimates of Conspiratorial Beliefs and Political Misperceptions", Political Analysis, (2023) vol.31: 575–590, DOI: 10.1017/pan.2022.2.

2-R. Bajo-Buenestado and M. A. Borrella-Mas, "The Heterogeneous Tax Pass-Through Under Different Vertical Relationships", The Economic Journal, (2022), vol.132(July): 1684–1708, DOI: 10.1093/ej/ueac007.

Materials: `https://osf.io/ncje7/`

Bogota Replication Games:

1-S. Hill and M. E. Roberts, "Acquiescence Bias Inflates Estimates of Conspiratorial Beliefs and Political Misperceptions", Political Analysis, (2023) vol.31: 575–590, DOI: 10.1017/pan.2022.2.

2-M. Comola and S. Prina, "The Interplay Among Savings Accounts and Network-Based Financial Arrangements: Evidence from a Field Experiment ", The Economic Journal, (2023), vol.133(January): 516–535, DOI: 10.1093/ej/ueac053.

Materials: `https://osf.io/hx67q/`

Tilburg Replication Games:

1-N. Lee, "Do Policy Makers Listen to Experts? Evidence from a National Survey of Local and State Policy Makers", American Political Science Review, (2022), vol.116(2): 677-688, DOI: 10.1017/S0003055421000800.

2-S. B. Holt and K. Vinopal, "Examining Inequality in the Time Cost of Waiting", Nature Human Behaviour, (2023), vol.7: 545–555, DOI: 10.1038/s41562-023-01524-w.

Materials: `https://osf.io/dqw5y/`

Virtual Replication Games: Europe (Part I)

1-P. Ager et al., "How the Other Half Died: Immigration and Mortality in U.S. Cities", The Review of Economic Studies, (2024), vol.91(1): 1–44, DOI: 10.1093/restud/rdad035

2-S. Herskowitz, "Gambling, Saving, and Lumpy Liquidity Needs", American Economic Journal: Applied Economics, (2021), vol.13(1): 72–104, DOI: 10.1257/app.20180177.

Materials: `https://osf.io/tcn7k/`

Virtual Replication Games: North America (Part II)

3-S. Herskowitz, “Gambling, Saving, and Lumpy Liquidity Needs”, American Economic Journal: Applied Economics, (2021), vol.13(1): 72–104, DOI: 10.1257/app.20180177.

4-A. G. de Zavala et al., “Mindful-Gratitude Practice Reduces Prejudice at High Levels of Collective Narcissism,” Psychological Science, (2024), vol.35(2): 137-149, DOI: 10.1177/09567976231220902.

Materials: `https://osf.io/67925/`

AI Training

Researchers took part in a 1-hour long training on the usage of ChatGPT. This training was mandatory for the researchers in AI-assisted and AI-led groups.

Recordings and materials are publicly available here: `https://osf.io/sz2g8/`.

The training included the following topics:

1) Introduction, Overview of ChatGPT, and Access

- Introduction to the capabilities of ChatGPT and its applications in reproducing scientific studies, coding, and data analysis.
- Instructions on accessing ChatGPT, creating an account, and accessing the Institute for Replication workspace/team subscription.
- Explanation of subscription tiers, model capabilities, limitations on message usage, and privacy settings.

2) Interaction with ChatGPT

- Techniques for optimizing prompts and ChatGPT’s responses, such as providing contextual information.
- Strategies to manage randomness in outputs or when the model gets ”stuck,” such as opening new chats and regenerating answers.

3) Sharing Chats

- Information on how to generate shareable links to sessions and manage privacy, including restrictions on who can access shared chats.

- Explanation on how to save chats as a webpage when the chat cannot be shared as a link (e.g., when the chat includes images).

4) Coding Assistance

- Explanation of how ChatGPT can assist with coding, including practical examples such as writing code for converting data formats (e.g., R's .rds to Stata's .dta) and debugging code.

5) File and Image Upload

- Introduction to ChatGPT's ability to process uploaded files.
- Overview of supported file types (e.g., PDFs, Word documents, CSVs, Excel files) and limitations regarding file size.
- Example of uploading an academic article to inquire about research questions, identification strategy, and robustness checks.
- Explanation of the potential benefits of uploading an image of a results table/figure instead of only the PDF file.
- Example of uploading an image of a results table from a study and inquiring about it.
- The image upload was not mentioned or demonstrated during the Toronto event.

6) Conducting Data Analysis Using ChatGPT

- Introduction to using ChatGPT's Data Analysis Module for executing Python code and performing data analysis.
- Example of uploading a replication package of an article and replicating regression analyses using the Python module.
- AI-led teams were instructed to first attempt to run the authors' codes/scripts using the data analysis capabilities of ChatGPT. If this analysis failed, teams were instructed to run the code in their local environment by following instructions provided by ChatGPT, as introduced in the Coding Assistance example.

7) ChatGPT API

- Explanation of the ChatGPT API for automating repetitive tasks and integrating AI capabilities into code.
- Example code shown for connecting to the ChatGPT API in R.

8) Customizing ChatGPT

- Information on setting up personalized models with custom instructions for specific needs.
- Mention of ChatGPT's memory feature that retains information across sessions, and how information that should not be retained can be deleted. The ChatGPT memory feature was not mentioned during the Toronto training session.

9) Explanation of Differences Among ChatGPT Models

- Differences between ChatGPT 4 and 4o were first discussed during the Sheffield training event.
- Introduction of GPT-o1-preview and GPT-o1-mini models was first provided during the Bogota event.
- Capabilities of ChatGPT 4o with canvas were introduced during the last event.

ChatGPT Models

Researchers in the AI-assisted and AI-led groups were provided with access to ChatGPT Team. Table S1 presents an overview of the ChatGPT models available to these researchers during each event. Table S2 provides details about the capabilities of these models. Throughout all events, researchers had access to the main flagship model, GPT-4, and/or GPT-4o. These models were capable of processing files, equipped with a Python environment for interpreting code and conducting data analysis, and had internet access.

The file upload was limited to maximum 512MB per file, and further limited to 2 million tokens for text files, approximately 50MB for CSV files and spreadsheets, 20MB per image for images. A user file size is capped at 10GB and organization at 100GB (*55*). However, the practical limitations based on the Python environment's capabilities were likely lower.

Games	Date	Training Date	Image*	ChatGPT versions available					
				3.5	4	4o	4o-mini	o1-preview / o1-mini	4o with canvas
Toronto	Feb 20	Feb 14	No	Yes	Yes				
Ottawa	May 3	Apr 26	Yes	Yes	Yes				
Sheffield	Jun 17	Jun 12	Yes	Yes	Yes	Yes			
Cornell	Aug 12	Jul 31	Yes		Yes	Yes	Yes		
Bogota	Oct 4	Sep 23**	Yes		Yes	Yes	Yes	Yes	Yes***
Tilburg	Oct 18	Sep 30	Yes		Yes	Yes	Yes	Yes	Yes***
Virtual	Nov 22	Nov 8	Yes		Yes	Yes	Yes	Yes	Yes

Table S1: ChatGPT models available by training

* Image upload trained as part of the pre-games training and screenshots of relevant results from the studies provided to researchers

** Training using recording of the Cornell training + o1-preview model slide added to presentation

*** While GPT-4o with canvas was available for the Bogota and Tilburg events, it was not mentioned during the training.

Only researchers in the Bogota, Tilburg, and the virtual only event had access to the GPT-o1-preview and GPT-o1-mini models. These models were trained using reinforcement learning to perform complex reasoning and, unlike the 4/4o models, can produce an internal chain of thought before responding to users. (*56*)

Usage limits for certain models were applied by OpenAI. During the Toronto and Ottawa events, these limits were explicitly stated, with the Team subscription limit set at 100 messages per three hours per user. Researchers were instructed to collaborate with their teammates if the limit was reached or use the unlimited GPT-3.5 model. For the remaining events, usage limits for the GPT-4/4o models were no longer explicitly mentioned by OpenAI but were likely higher. The GPT-o1-preview model was limited to 50 queries per week, while GPT-o1-mini was limited to 50 queries per day.

Model	Date Introduced	File Upload	Python Code Interpreter	Web Browsing	"Thinking"
GPT-3.5	Before 1st event	No	No	No	No
GPT-4	Before 1st event	Yes	Yes	Yes	No
GPT-4o	May 13 [1]	Yes	Yes	Yes	No
GPT-4o-mini	July 18 [2]	Yes*	Yes*	Yes*	No
GPT-o1-preview	September 12 [3]	No	No	No	Yes
GPT-o1-mini	September 12 [3]	No	No	No	Yes
GPT-4o with canvas	October 3 [4]	Yes	Yes	No	No

Table S2: ChatGPT capabilities

* While 4o-mini supported these functions at the time of the last training it did not necessarily at the time of introduction.

[1] `https://openai.com/index/hello-gpt-4o/`

[2] `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`

[3] `https://openai.com/index/introducing-openai-o1-preview/`

[4] `https://openai.com/index/introducing-canvas/`

Coding Errors and Data Irregularities

We define coding errors as minor or major depending on whether the coding error could, in theory, have an impact on the claims tested. AB, JA and DM discussed all errors uncovered and classified coding errors as major or minor. Coding errors uncovered range from minor errors such as missing packages/paths or versioning issues to major coding errors such as miscoding the dependent variable or main independent variable and conducting a many-to-many merge instead of a many-to-one merge.

In what follows, we provide concrete examples of major coding errors and data irregularities. In the article entitled "Arrests and Convictions but Not Sentence Length Deter Terrorism in 28 European Union Member States", one of the major coding errors is in the coding of the dependent variable. The authors state in the article that the terrorism rate used as their dependent variable is the inverse hyperbolic sine (IHS) of the per capita rate of terrorist attacks. But the code reveals that the dependent variable takes impossible values and is thus not the IHS of the per capita rate of terrorist attacks. For instance, countries with zero terror attacks are assigned strictly positive values, which is not possible. Another major coding error is that some European countries were imputed having zero terror attacks because of joining the European Union during the sample time period. This coding error is due to the terrorism dataset only covering European Union countries and the authors assigning zero values instead of missing information for these countries. There is an editor's note for this article at Nature Human Behaviour. The note was released as a result of a Matters Arising submission by one of our reproducers. The Matters Arising is revised and resubmitted.

In the article "Sorting or Steering: The Effects of Housing Discrimination on Neighborhood Choice", one of the major coding errors involved assigning a value of zero for the variable "of color" to both individuals identified as 'white' and as 'other' in the raw data. A major data irregularity is the inclusion of fixed effects for the string variable 'city'. The raw variable is case sensitive and has many spelling mistakes. A comment detailing these errors is revised and resubmitted at the Journal of Political Economy.

In the main article, we document that AI-led groups identified significantly less coding errors and data irregularities. AI-led groups likely uncovered fewer major coding errors due to the nuanced

and contextual nature of these errors. It is plausible that AI-led groups struggled to identify technically correct but conceptually flawed code errors. These errors, such as a many-to-many merge instead of a many-to-one merge, produce duplicate entries without causing a runtime error. While the code executes without issues, the underlying conceptual mistake leads to incorrect data handling. This type of error is particularly challenging for AI to detect, as it requires an understanding of the conceptual intent behind the code rather than just its syntactic correctness.

More generally, many coding mistakes involve subtle misapplications of statistical transformations, such as assigning incorrect values or mishandling missing data, which often require domain expertise and a deep understanding of the data's structure. AI tools, while efficient at automating tasks, may struggle with interpreting complex logical relationships, ambiguous data definitions, or recognizing implausible outcomes without explicit programming. In contrast, human-led groups are better equipped to identify errors that hinge on contextual reasoning, such as the incorrect coding of dependent variables or misassignments due to case-sensitive inconsistencies in datasets.

Robustness Checks

We propose four different binary measures which we believe qualify a good robustness check: (i) clarity (not vague) regarding purpose and execution; (ii) feasible, (iii) not previously done by the original author(s); and (iv) focuses on the validity of the empirical strategy. In addition, we classify any corrections to coding errors and rerunning the script would be considered a "good" robustness check, although not checking the above list. One of AB, JA and DM reviewed each robustness check based on the above criteria. In the event that at least one of the above categories is hard to classify, we discussed and classified together.

Clarity (not vague) regarding purpose and execution: It is possible that teams of replicators will not adequately describe their robustness check. This could be due to ChatGPT not sufficiently describing what they are doing, or, from their own explanation. An example of a vague robustness check would be "adding control variables." In contrast, a clear robustness check would be to precisely document which variable should be added as a control.

Feasible: Feasibility both corresponds to what could be done. In the former, teams who are able to perform robustness checks have performed a feasible robustness check. For those which cannot execute a robustness check, the question we ask is whether or not, with more time but the same

resources, if it could be done.

Not previously done by the original author(s): All teams of reproducers - independent of which type of team they are - have access to the original study, online appendix and the replication packages. All teams ought to be capable of verifying their recommended robustness check was not previously done. AB, JA and DM verified with each study whether the proposed robustness check were included in the article or appendix.

Validity: While robustness checks can serve multiple purposes, we view them as alternative specifications which test the main conclusion(s) of a study. A valid robustness check tests the reliability and stability of the results. Examples of invalid robustness checks include: using bad controls, misspecified models (bad instrument), etc.

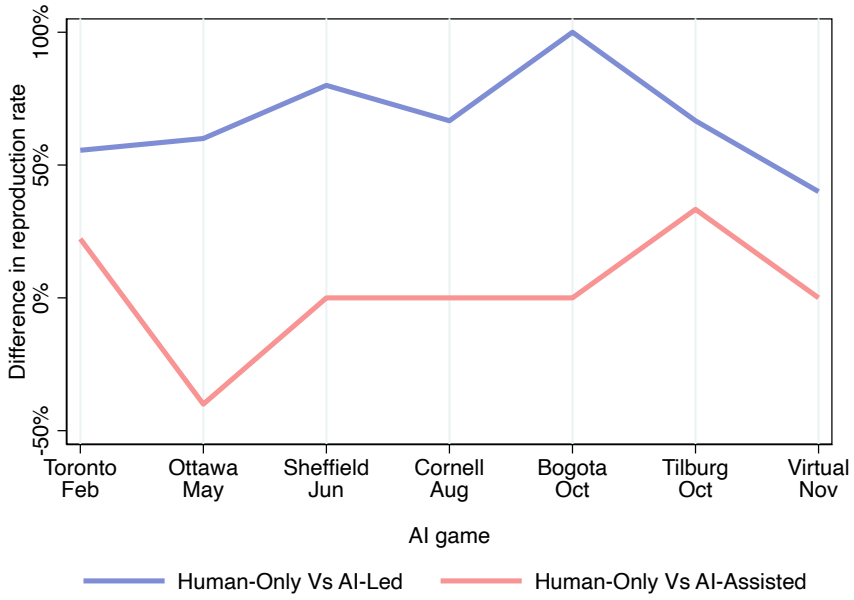

Figure S1: Difference in computational reproducibility rate across groups for each event.

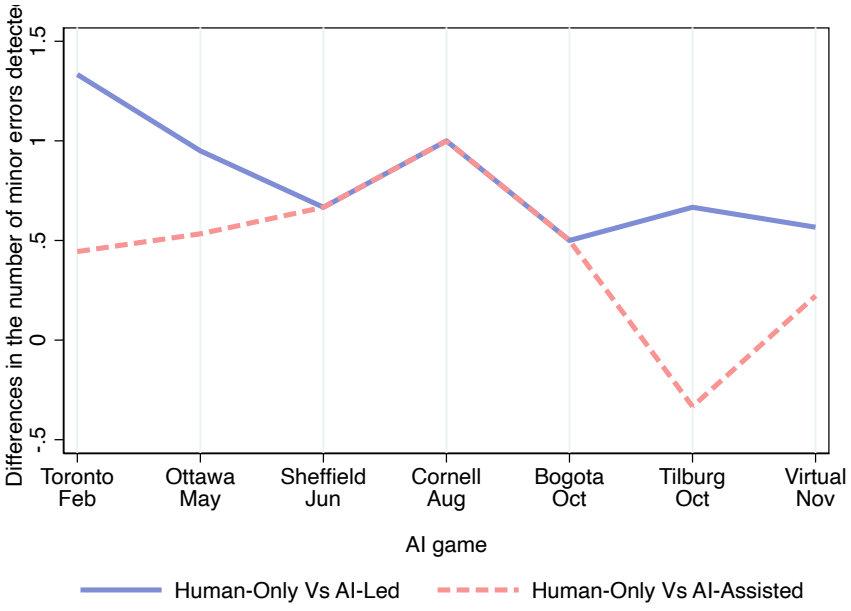

Figure S2: Differences in the number of minor errors detected across groups for each event.

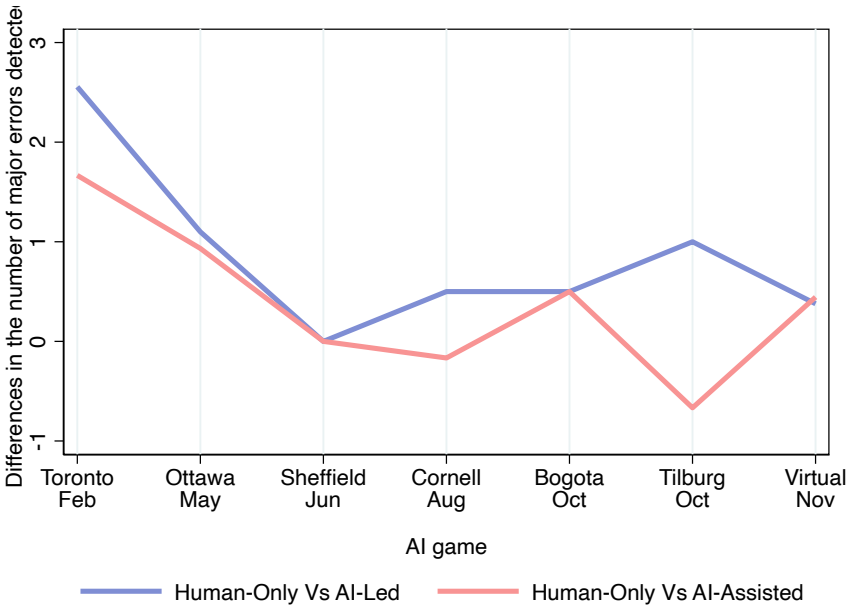

Figure S3: Differences in the number of major errors detected across groups for each event.

Table S3: Balance table of Human-Only, AI-Assisted, and AI-Led Metrics

Variable	Human-Only	AI-Assisted	AI-Led	Human-Only vs AI-Assisted	Human-Only vs AI-Led	AI-Assisted vs AI-Led
Software: R	0.545 (0.506)	0.543 (0.505)	0.600 (0.497)	0.003 (0.983)	-0.055 (0.655)	-0.057 (0.635)
Number of teammates	2.606 (0.496)	2.429 (0.655)	2.829 (0.568)	0.177 (0.214)	-0.223 (0.091)	-0.400 (0.008)
Attendance: In-Person	0.333 (0.479)	0.343 (0.482)	0.257 (0.443)	-0.010 (0.935)	0.076 (0.498)	0.086 (0.441)
Minimum academic level: Professor	0.091 (0.292)	0.086 (0.284)	0.086 (0.284)	0.005 (0.941)	0.005 (0.941)	-0.000 (1.000)
Minimum academic level: Postdoc	0.030 (0.174)	0.114 (0.323)	0.057 (0.236)	-0.084 (0.190)	-0.027 (0.597)	0.057 (0.400)
Minimum academic level: Researcher	0.152 (0.364)	0.171 (0.382)	0.029 (0.169)	-0.020 (0.827)	0.123 (0.076)	0.143 (0.047)
Minimum academic level: Student	0.727 (0.452)	0.629 (0.490)	0.829 (0.382)	0.099 (0.392)	-0.101 (0.321)	-0.200 (0.061)
Maximum academic level: Professor	0.576 (0.502)	0.514 (0.507)	0.686 (0.471)	0.061 (0.617)	-0.110 (0.355)	-0.171 (0.147)
Maximum academic level: Postdoc	0.152 (0.364)	0.257 (0.443)	0.143 (0.355)	-0.106 (0.289)	0.009 (0.921)	0.114 (0.238)
Maximum academic level: Researcher	0.091 (0.292)	0.057 (0.236)	0.000 (0.000)	0.034 (0.600)	0.091 (0.070)	0.057 (0.156)
Maximum academic level: Student	0.182 (0.392)	0.171 (0.382)	0.171 (0.382)	0.010 (0.912)	0.010 (0.912)	-0.000 (1.000)
Average years of coding experience	9.444 (4.781)	7.967 (2.927)	9.712 (3.230)	1.478 (0.127)	-0.267 (0.787)	-1.745 (0.021)
Min ChatGPT level: Never	0.281 (0.457)	0.143 (0.355)	0.286 (0.458)	0.138 (0.169)	-0.004 (0.968)	-0.143 (0.150)
Min ChatGPT level: Beginner	0.594 (0.499)	0.543 (0.505)	0.600 (0.497)	0.051 (0.680)	-0.006 (0.959)	-0.057 (0.635)
Min ChatGPT level: Intermediate	0.094 (0.296)	0.314 (0.471)	0.086 (0.284)	-0.221 (0.027)	0.008 (0.910)	0.229 (0.017)
Min ChatGPT level: Advanced	0.031 (0.177)	0.000 (0.000)	0.029 (0.169)	0.031 (0.299)	0.003 (0.950)	-0.029 (0.321)
Max ChatGPT level: Never	0.000 (0.000)	0.029 (0.169)	0.029 (0.169)	-0.029 (0.343)	-0.029 (0.343)	-0.000 (1.000)
Max ChatGPT level: Beginner	0.188 (0.397)	0.114 (0.323)	0.086 (0.284)	0.073 (0.409)	0.102 (0.229)	0.029 (0.695)
Max ChatGPT level: Intermediate	0.469 (0.507)	0.514 (0.507)	0.714 (0.458)	-0.046 (0.715)	-0.246 (0.041)	-0.200 (0.088)
Max ChatGPT level: Advanced	0.344 (0.483)	0.343 (0.482)	0.171 (0.382)	0.001 (0.994)	0.172 (0.109)	0.171 (0.104)

Note: Standard errors in parentheses for individual branches (Human-only, AI-Assisted, and AI-Led); p-values in parentheses for branch comparisons (Human-Only Vs AI-Assisted, Human-Only Vs AI-Led, and AI-Assisted Vs AI-Led).

Table S4: Results from Logit & Poisson regressions predicting reproduction outcomes

	(1) Reproduction	(2) Minor errors	(3) Major errors	(4) Two good robustness	(5) Ran one robustness	(6) Ran two robustness
main						
AI-Assisted	-0.609	-0.454**	-0.667**	-0.332	-0.725	-0.015
	(1.258)	(0.189)	(0.261)	(0.825)	(1.458)	(0.689)
	[-3.075; 1.858]	[-0.823; -0.084]	[-1.179; -0.155]	[-1.949; 1.284]	[-3.582; 2.133]	[-1.366; 1.335]
AI-Led	-4.901***	-1.177***	-1.802***	-2.071**	-2.750**	-1.395**
	(1.506)	(0.200)	(0.365)	(0.865)	(1.144)	(0.637)
	[-7.853; -1.949]	[-1.568; -0.786]	[-2.517; -1.087]	[-3.767; -0.375]	[-4.992; -0.507]	[-2.643; -0.147]
Model	Logit	Poisson	Poisson	Logit	Logit	Logit
Controls	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.713	1.000	0.893	0.699	0.779	0.660
p-val (AI-Assisted vs. AI-Led)	0.001	0.003	0.004	0.030	0.066	0.035
Observations	94	98	84	73	86	97

Note: Standard errors in parentheses, confidence intervals in brackets; human-only branch omitted; the model for One good robustness is not included due to unsuficient observations, preventing it from converging. Marginal effects reported for Logit models.

Controls include number of teammates; game-software, skill, and attendance fixed effects.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S5: Results from OLS regressions predicting reproduction outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Reproduction	Minor errors	Major errors	One good robustness	Two good robustness	Ran one robustness	Ran two robustness
Branch: AI-Assisted	-0.018	-0.487*	-0.646**	-0.009	-0.014	-0.032	-0.009
	(0.063)	(0.270)	(0.254)	(0.027)	(0.103)	(0.061)	(0.113)
	[-0.144; 0.107]	[-1.025; 0.051]	[-1.153; -0.139]	[-0.063; 0.046]	[-0.220; 0.191]	[-0.155; 0.090]	[-0.233; 0.216]
Branch: AI-Led	-0.593***	-1.050***	-1.136***	-0.167**	-0.250**	-0.323***	-0.290**
	(0.090)	(0.258)	(0.235)	(0.068)	(0.107)	(0.098)	(0.126)
	[-0.773; -0.413]	[-1.565; -0.536]	[-1.604; -0.667]	[-0.302; -0.031]	[-0.463; -0.037]	[-0.518; -0.127]	[-0.540; -0.040]
Number of teammates	0.052	0.344	0.224	-0.026	0.091	-0.038	0.059
	(0.069)	(0.208)	(0.196)	(0.050)	(0.083)	(0.077)	(0.094)
	[-0.085; 0.189]	[-0.070; 0.758]	[-0.168; 0.615]	[-0.125; 0.073]	[-0.073; 0.256]	[-0.192; 0.116]	[-0.128; 0.245]
Game: Ottawa	-0.086	-1.248	-0.243	-0.049	0.098	-0.351*	-0.069
	(0.155)	(0.755)	(0.460)	(0.172)	(0.217)	(0.177)	(0.205)
	[-0.394; 0.222]	[-2.752; 0.256]	[-1.159; 0.673]	[-0.392; 0.293]	[-0.333; 0.530]	[-0.704; 0.002]	[-0.478; 0.340]
Game: Sheffield	-0.180	-1.604**	-0.653	0.159	-0.006	-0.336	-0.010
	(0.259)	(0.644)	(0.464)	(0.136)	(0.400)	(0.310)	(0.360)
	[-0.696; 0.336]	[-2.886; -0.322]	[-1.577; 0.271]	[-0.111; 0.430]	[-0.803; 0.792]	[-0.954; 0.281]	[-0.727; 0.707]
Game: Cornell	0.276	-1.547***	-1.075**	0.061	0.317	0.055	0.285
	(0.190)	(0.540)	(0.499)	(0.120)	(0.201)	(0.180)	(0.235)
	[-0.103; 0.655]	[-2.622; -0.471]	[-2.069; -0.081]	[-0.178; 0.299]	[-0.083; 0.717]	[-0.304; 0.415]	[-0.183; 0.754]
Game: Bogota	0.014	-1.074	-1.465	0.016	0.071	-0.189	-0.170
	(0.175)	(1.000)	(1.021)	(0.136)	(0.350)	(0.178)	(0.354)
	[-0.334; 0.362]	[-3.065; 0.917]	[-3.498; 0.568]	[-0.255; 0.287]	[-0.626; 0.768]	[-0.543; 0.165]	[-0.875; 0.535]
Game: Tilburg	0.231	-2.185***	0.532	0.067	0.398**	-0.076	0.230
	(0.173)	(0.697)	(0.720)	(0.112)	(0.168)	(0.175)	(0.184)
	[-0.114; 0.575]	[-3.573; -0.798]	[-0.901; 1.965]	[-0.155; 0.289]	[0.063; 0.734]	[-0.424; 0.272]	[-0.136; 0.597]
Game: Virtual Europe	0.004	-1.976***	-0.818	0.098	0.351**	0.092	0.245
	(0.161)	(0.623)	(0.587)	(0.110)	(0.174)	(0.128)	(0.250)
	[-0.316; 0.324]	[-3.216; -0.736]	[-1.988; 0.351]	[-0.120; 0.317]	[0.005; 0.696]	[-0.162; 0.347]	[-0.254; 0.743]
Game: Virtual North America	0.123	-1.545***	-0.583	0.088	0.327**	-0.130	0.092
	(0.180)	(0.469)	(0.458)	(0.112)	(0.163)	(0.179)	(0.205)
	[-0.235; 0.481]	[-2.478; -0.612]	[-1.495; 0.329]	[-0.135; 0.311]	[0.004; 0.651]	[-0.487; 0.227]	[-0.315; 0.500]
Software: R	-0.169	0.915	0.249	0.006	0.118	-0.014	0.096
	(0.154)	(0.621)	(0.513)	(0.121)	(0.183)	(0.123)	(0.184)
	[-0.476; 0.138]	[-0.322; 2.152]	[-0.772; 1.271]	[-0.236; 0.248]	[-0.246; 0.481]	[-0.260; 0.231]	[-0.270; 0.463]
Game: Ottawa × Software: R	-0.281	-1.622**	-0.306	-0.094	-0.066	-0.216	-0.352
	(0.285)	(0.741)	(0.647)	(0.250)	(0.307)	(0.324)	(0.353)
	[-0.849; 0.287]	[-3.098; -0.147]	[-1.594; 0.982]	[-0.592; 0.405]	[-0.677; 0.545]	[-0.861; 0.429]	[-1.055; 0.352]
Game: Sheffield × Software: R	0.322	-1.355**	-0.751	-0.276	-0.260	0.178	-0.321
	(0.316)	(0.632)	(0.621)	(0.200)	(0.458)	(0.341)	(0.430)
	[-0.307; 0.952]	[-2.614; -0.096]	[-1.987; 0.485]	[-0.674; 0.122]	[-1.172; 0.651]	[-0.500; 0.856]	[-1.178; 0.535]
Game: Cornell × Software: R	-0.265	-1.309**	0.424	-0.012	-0.115	-0.449*	-0.557*
	(0.244)	(0.614)	(0.742)	(0.151)	(0.239)	(0.227)	(0.303)
	[-0.750; 0.219]	[-2.532; -0.086]	[-1.054; 1.902]	[-0.313; 0.289]	[-0.592; 0.361]	[-0.901; 0.002]	[-1.160; 0.046]
Game: Bogota × Software: R	0.112	-1.867*	0.727	0.010	0.103	-0.054	0.044
	(0.319)	(1.100)	(1.032)	(0.165)	(0.384)	(0.259)	(0.420)
	[-0.524; 0.747]	[-4.058; 0.324]	[-1.329; 2.783]	[-0.319; 0.339]	[-0.661; 0.867]	[-0.570; 0.462]	[-0.791; 0.880]
Game: Tilburg × Software: R	-0.364	-0.235	-1.553*	-0.026	-0.697**	-0.250	-0.915***
	(0.252)	(0.802)	(0.898)	(0.134)	(0.294)	(0.310)	(0.270)
	[-0.866; 0.137]	[-1.832; 1.362]	[-3.340; 0.235]	[-0.294; 0.241]	[-1.283; -0.110]	[-0.867; 0.367]	[-1.453; -0.377]
Game: Virtual Europe × Software: R	0.234	-0.837	-0.597	-0.024	-0.424	-0.111	-0.287
	(0.222)	(0.774)	(0.745)	(0.140)	(0.287)	(0.189)	(0.341)
	[-0.208; 0.676]	[-2.379; 0.705]	[-2.080; 0.887]	[-0.303; 0.255]	[-0.995; 0.147]	[-0.486; 0.265]	[-0.966; 0.392]
Game: Virtual North America × Software: R	0.076	-1.145	-0.216	0.002	-0.312	0.110	-0.210
	(0.285)	(0.723)	(0.836)	(0.138)	(0.316)	(0.275)	(0.377)
	[-0.491; 0.643]	[-2.585; 0.294]	[-1.881; 1.449]	[-0.273; 0.277]	[-0.941; 0.317]	[-0.437; 0.658]	[-0.961; 0.540]
Maximum academic level: Researcher	0.148	-1.548**	0.336	-0.015	-0.213	0.105	-0.053
	(0.180)	(0.628)	(1.559)	(0.091)	(0.195)	(0.159)	(0.219)
	[-0.211; 0.506]	[-2.799; -0.298]	[-2.769; 3.440]	[-0.197; 0.168]	[-0.601; 0.174]	[-0.212; 0.421]	[-0.490; 0.384]
Maximum academic level: Postdoc	0.110	0.078	0.275	0.032	-0.090	0.314**	0.196
	(0.177)	(0.337)	(0.318)	(0.082)	(0.145)	(0.146)	(0.172)
	[-0.243; 0.463]	[-0.594; 0.750]	[-0.357; 0.908]	[-0.131; 0.195]	[-0.379; 0.199]	[0.022; 0.605]	[-0.147; 0.538]
Maximum academic level: Professor	0.030	0.008	0.340	-0.043	-0.165	0.107	-0.008
	(0.140)	(0.245)	(0.262)	(0.090)	(0.128)	(0.145)	(0.147)
	[-0.248; 0.309]	[-0.480; 0.495]	[-0.183; 0.862]	[-0.223; 0.136]	[-0.419; 0.089]	[-0.181; 0.395]	[-0.300; 0.283]
Minimum academic level: Researcher	-0.140	-0.126	0.285	-0.012	0.184	0.078	0.269*
	(0.091)	(0.345)	(0.519)	(0.066)	(0.116)	(0.108)	(0.137)
	[-0.322; 0.042]	[-0.813; 0.560]	[-0.748; 1.318]	[-0.142; 0.119]	[-0.047; 0.415]	[-0.137; 0.292]	[-0.004; 0.541]
Minimum academic level: Postdoc	-0.080	0.089	-0.150	-0.127	-0.082	0.033	0.005
	(0.214)	(0.851)	(0.435)	(0.134)	(0.183)	(0.123)	(0.195)
	[-0.506; 0.346]	[-1.604; 1.783]	[-1.015; 0.715]	[-0.394; 0.141]	[-0.447; 0.284]	[-0.213; 0.279]	[-0.383; 0.394]
Minimum academic level: Professor	-0.094	0.697	0.535	0.001	-0.081	0.006	-0.076
	(0.150)	(0.431)	(0.414)	(0.051)	(0.187)	(0.141)	(0.223)
	[-0.393; 0.204]	[-0.161; 1.555]	[-0.290; 1.359]	[-0.101; 0.102]	[-0.452; 0.291]	[-0.274; 0.287]	[-0.520; 0.369]
Attendance: In-Person	-0.124	0.327	0.205	-0.012	-0.072	0.227*	0.196
	(0.120)	(0.369)	(0.245)	(0.085)	(0.121)	(0.120)	(0.127)
	[-0.362; 0.115]	[-0.409; 1.062]	[-0.282; 0.693]	[-0.181; 0.156]	[-0.312; 0.168]	[-0.013; 0.467]	[-0.057; 0.450]
Mean of dep. var	0.738	0.951	0.728	0.942	0.786	0.816	0.680
p-val (AI-Assisted vs. AI-Led)	0.000	0.022	0.018	0.023	0.036	0.004	0.019
Observations	103	103	103	103	103	103	103

Note: Standard errors in parentheses, confidence intervals in brackets; human-only branch omitted.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S6: Results from OLS regressions predicting reproduction outcomes for Stata teams

	(1) Reproduction	(2) Minor errors	(3) Major errors	(4) One good robustness	(5) Two good robustness	(6) Ran one robustness	(7) Ran two robustness
AI-Assisted	0.002	-0.886**	-0.990***	0.009	0.159	0.048	0.260*
	(0.075)	(0.375)	(0.350)	(0.047)	(0.111)	(0.068)	(0.135)
	[-0.153; 0.157]	[-1.655; -0.116]	[-1.708; -0.272]	[-0.087; 0.105]	[-0.069; 0.387]	[-0.092; 0.188]	[-0.016; 0.537]
AI-Led	-0.577***	-1.227**	-1.679***	-0.115	-0.233	-0.406**	-0.231
	(0.155)	(0.503)	(0.392)	(0.102)	(0.182)	(0.154)	(0.227)
	[-0.894; -0.260]	[-2.259; -0.195]	[-2.483; -0.876]	[-0.324; 0.094]	[-0.607; 0.140]	[-0.723; -0.090]	[-0.696; 0.235]
Mean of dep. var	0.844	0.889	0.844	0.956	0.844	0.867	0.778
p-val (AI-Assisted vs. AI-Led)	0.001	0.289	0.052	0.195	0.024	0.006	0.014
observations	45	45	45	45	45	45	45

Note: Standard errors in parentheses, Confidence intervals in brackets; human-only branch omitted; sample restricted to papers for which the replication package is mainly in Stata

Controls include number of teammates; game, skill, and attendance fixed effects.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S7: Results from OLS regressions predicting reproduction outcomes for R teams

	(1) Reproduction	(2) Minor errors	(3) Major errors	(4) One good robustness	(5) Two good robustness	(6) Ran one robustness	(7) Ran two robustness
AI-Assisted	-0.002	-0.185	0.049	-0.040	-0.231	-0.101	-0.285
	(0.144)	(0.305)	(0.321)	(0.059)	(0.189)	(0.125)	(0.203)
	[-0.294; 0.289]	[-0.802; 0.432]	[-0.600; 0.698]	[-0.159; 0.079]	[-0.612; 0.150]	[-0.353; 0.151]	[-0.694; 0.125]
AI-Led	-0.612***	-0.937***	-0.661***	-0.165*	-0.266*	-0.311**	-0.362**
	(0.124)	(0.278)	(0.241)	(0.087)	(0.152)	(0.136)	(0.177)
	[-0.863; -0.361]	[-1.498; -0.375]	[-1.148; -0.175]	[-0.340; 0.010]	[-0.573; 0.040]	[-0.587; -0.036]	[-0.720; -0.003]
Mean of dep. var	0.655	1.000	0.638	0.931	0.741	0.776	0.603
p-val (AI-Assisted vs. AI-Led)	0.002	0.016	0.029	0.163	0.837	0.195	0.681
observations	58	58	58	58	58	58	58

Note: Standard errors in parentheses, Confidence intervals in brackets; human-only branch omitted; sample restricted to papers for which the replication package is mainly in R

Controls include number of teammates; game, skill, and attendance fixed effects.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table S8: Comparison of AI-Assisted and AI-Led Metrics by Experience Level

Variable	AI-Assisted high experience (n=12)	AI-Assisted low/medium experience (n=23)	AI-Led high experience (n=6)	AI-Led low/medium experience (n=29)
Reproduction	0.833 (0.389)	0.957 (0.209)	0.167 (0.408)	0.414 (0.501)
Minutes to reproduction	104.8 (89.5)	86.7 (89.2)	116.0 (-)	180.0 (66.0)
Number of minor errors	1.250 (1.288)	0.783 (1.536)	0.500 (0.837)	0.517 (0.949)
Minutes to first minor error	87.6 (44.5)	61.9 (47.5)	245.0 (106.9)	114.5 (89.0)
Number of major errors	0.833 (0.937)	0.522 (0.947)	0.500 (0.837)	0.172 (0.384)
Minutes to first major error	105.7 (36.4)	146.1 (59.5)	362.0 (-)	163.3 (68.3)
At least one good robustness check	1.000 (0.000)	1.000 (0.000)	0.667 (0.516)	0.862 (0.351)
At least two good robustness checks	0.833 (0.389)	0.870 (0.344)	0.500 (0.548)	0.655 (0.484)
Ran at least one good robustness check	0.917 (0.289)	0.957 (0.209)	0.000 (0.000)	0.690 (0.471)
Ran At least two good robustness checks	0.750 (0.452)	0.826 (0.388)	0.000 (0.000)	0.552 (0.506)

Note: Standard errors in parentheses; experience is categorized based on the level of experience with ChatGPT of the most advanced team member (prior to AI training). High Experience corresponds to Advanced, and Low/Intermediate Experience corresponds to Never, Beginner, and Intermediate levels.

Table S9: Comparison of Key Metrics by Prompt Levels within AI-Assisted Branch

Variable	Above median (n=17)	Below/equal to median (n=18)
Reproduction	1.000 (0.000)	0.833 (0.383)
Minutes to reproduction	118.5 (111.8)	62.7 (34.5)
Number of minor errors	0.824 (1.131)	1.056 (1.731)
Minutes to first minor error	86.6 (47.2)	62.8 (45.6)
Number of major errors	0.529 (0.943)	0.722 (0.958)
Minutes to first major error	153.7 (64.1)	110.1 (37.8)
At least one good robustness check	1.000 (0.000)	1.000 (0.000)
At least two good robustness checks	0.824 (0.393)	0.889 (0.323)
Ran at least one good robustness check	0.941 (0.243)	0.944 (0.236)
Ran at least two good robustness checks	0.765 (0.437)	0.833 (0.383)

Note: Standard errors in parentheses; Groups are defined based on the median number of prompts (19) in the AI-Assisted sample.