PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES

AI-assisted teams outperform AI-led teams but not human-only teams in assessing research reproducibility in quantitative social science

Abel Brodeur[a,b,1], David Valenta[a,b], Alexandru Marcoci[c], Juan P. Aparicio[a,b], Derek Mikola[a,b], Bruno Barbarioli[a,b], Rohan Alexander[d], Lachlan Deer[e], Tom Stafford[f,g], Lars Vilhuber[h], Gunther Bensch[i], Fabio Motoki[j,k], Mohamed Abdelhady[l], Yousra Abdelmoula[m], Ghina Abdul Baki[a,b], Tomás Aguirre[n], Sriraj Aiyer[o], Shumi Akhtar[p], Farida Akhtar[q], Melle R. Albada[r], Micah Altman[s], David Angenendt[t,u], Zahra Arjmandi Lari[v], Jorge Armando De León Tejada[w], David Rodriguez Arana[x], Igor Asanov[y], Anastasiya-Mariya Noha[y], Rebecca Ashong[z], Tobias Auer[aa], Francisco J. Bahamonde-Birke[bb], Bradley J. Baker[cc], Söhnke M. Bartram[dd,ee], Dongqi Bao[ff], Lucija Batinovic[gg], Tommaso Batistoni[hh], Monica Beeder[ii], Louis-Philippe Beland[l], Carsten Gero Bienz[jj], Christ Billy Aryanto[kk], Cylcia Bolibaugh[ll], Carl Bonander[mm,nn], Ramiro Bravo[oo], Egor Bronnikov[pp,qq], Stephan Bruns[rr,ss,tt], Nino Buliskeria[uu], Sara Caicedo-Silva[vv], Andrea Calef[ww], Juan Sebastian Cano Arias[x], Gustavo A. Castillo Alvarez[xx], Solomon Caulker[yy], Simonas Cepenas[zz], Arthur Chatton[aaa,bbb], Zirou Chen[ccc], Ngozi Chioma Ewurum[ddd], Anda-Bianca Ciocîrlan[f], Felix J. Clouth[bb], Jason Collins[eee], Nikolai Cook[fff], Cesar Cornejo[ggg,hhh], João Craveiro[f], Jonathan Créchet[iii], Jing Cui[jjj,kkk], Niveditha Chalil Vayalabron[lll], Christian Czymara[mmm], Carlos Daniel Bermúdez Jaramillo[nnn], Hannes Datta[ooo], Lien Denoo[ppp], Arshia Dhaliwal[l], Nency Dhameja[qqq], Elodie Djemai[rrr], Erwan Dujeancourt[sss,ttt], Uğurcan Dündar[uuu], Thibaut Duprey[vvv], Yasmine Eissa[www], Youssef El Fassi[xxx], Ismail El Fassi[yyy], Keaton Ellis[zzz], Ali Elminejad[uu], Mahmoud Elsherif[aaaa,bbbb], Aysil Emirmahmutoglu[cccc], Giulian Etingin-Frati[dddd], Emeka Eze[eeee], Jan Fabian Dollbaum[ffff], Jan Feld[gggg], Andres Felipe Rengifo Jaramillo[hhhh,iiii], Guidon Fenig[a], Victoria Fernandes[vv,jjjj], Lenka Fiala[a,b,kkkk], Lukas Fink[llll], Mojtaba Firouzjaeiangalougah[mmmm], Sara Fish[nnnn], Jack Fitzgerald[oooo,pppp], Rachel Forshaw[qqqq], Alexandre Fortier-Chouinard[rrrr], Louis Fréget[ssss], Joris Frese[tttt], Jacopo Gabani[uuuu,vvvv], Sebastian Gallegos[wwww], Max C. Gamill[xxxx], Attila Gáspár[yyyy,zzzz], Romain Gauriot[aaaaa], Evelina Gavrilova[cccc], Diogo Geraldes[bbbbb,ccccc,ddddd], Giulio Giacomo Cantone[eeeee], Grant Gibson[fffff,ggggg], Dirk Goldschmitt[f], Amélie Gourdon-Kanhukamwe[hhhhh], Andrea Gregor de Varda[iiiii], Idaliya Grigoryeva[jjjjj], Alexi Gugushvili[kkkkk], Aaron H. A. Fletcher[lllll], Florian Habermann[mmmmm,nnnnn], Márton Hablicsek[ooooo], Joanne Haddad[ppppp], Jonathan D. Hall[qqqqq], Olle Hammar[ttt,rrrrr], Malek Hassouneh[sssss], Carina I. Hausladen[ttttt], Sophie C. F. Hendrikse[bb], Matthew Hepplewhite[uuuuu], Anson T. Y. Ho[vvvvv], Senan Hogan-Hennessy[h], Elliot Howley[wwwww], Gaoyang Huang[xxxxx,ff], Héloïse Hulstaert[rr,yyyyy], Zlatomira G. Ilchovska[zzzzz,aaaaaa,wwwwww], Paola Jaimes Santamaria[bbbbbb,cccccc], Niklas Jakobsson[mm], Joakim Jansson[ttt,dddddd], Ewa Jarosz[eeeeee], Hossein Jebeli[ffffff], Yanchen Jiang[nnnn], Hiba Junaid[gggggg,hhhhhh], Rohan Kalluraya[iiiiii], Sunny Karim[jjjjjj], Edmund Kelly[kkkkkk], Eva Kimel[zzzzz], Sorravich Kingsuwankul[llllll,pppp], Valentin Klotzbücher[mmmmmm,nnnnnn,oooooo], Daniel Krähmer[pppppp], Pijus Krūminas[zz], Nicholas Kruus[uuuuu], Essi Kujansuu[qqqqqq,rrrrrr], Christoph F. Kurz[ssssss], Stephan Küster[tttttt], Blake Lee-Whiting[uuuuuu], Felix Lewandowski[wwwwww], Tongzhe Li[vvvvvv], Ruoxi Li[wwwwwww], Dan Liu[xxxxxx], Jiacheng Liu[yyyyyy], Helix Lo[zzzzzz], Katharina Loter[bb], Felipe Macedo Dias[aaaaaaa], Christopher R. Madan[wwwwww], Nicolas Mäder[bbbbbbb], Marco Mandas[ccccccc], Cesar Mantilla[ddddddd], Jan Marcus[llll], Diego Marino Fages[eeeeeee], Xavier Martin[fffffff], Ryan McWay[ggggggg], Daniel Medina-Gaspar[hhhhhhh], Sisi Meng[iiiiiii], Lingyu Meng[jjjjjjj], Simon Merz[kkkkkkk], Alex P. Miller[lllllll], Thibault Mirabel[mmmmmmm], Dibya Deepta Mishra[nnnnnnn], Sumit Mishra[ooooooo], Belay W. Moges[ppppppp], Morteza Mohandes Mojarrad[qqqqqqq], Myra Mohnen[rrrrrrr], Louis-Philippe Morin[a], Lucija Muehlenbachs[sssssss,ttttttt], Gastón Mullin[uuuuuuu], Andreea Musulan[vvvvvvv,wwwwwww], Sara Muzzì[xxxxxxx,yyyyyyy], James A. C. Myers[zzzzzzz], Florian Neubauer[aaaaaaaa], Tuan Nguyen[rr], Ali Niazi[sssssss], Ardyn Nordstrom[bbbbbbbb], Bartłomiej Nowak[cccccccc], Daneal O'Habib[dddddddd], Tim Ölkers[eeeeeeee], Justin Ong[f], Valeria Orozco Castiblanco[ffffffff,gggggggg], Ömer Özak[hhhhhhhh], Ali I. Ozkes[iiiiiiii,jjjjjjjj], Mikael Paaso[kkkkkkkk], Shubham Pandey[llllllll], Varvara Papazoglou[lllll], Romeo Penheiro[mmmmmmmm], Linh Pham[nnnnnnnn], Ulrike Phieler[oooooooo], Peter Pütz[pppppppp], Quan Qi[qqqqqqqq], Jingyi Qiu[rrrrrrrr], Manuel T. Rein[bb], David A. Reinstein[ssssssss], Juuso Repo[tttttttt], Nicolas Rudolf[nnnnn], Shree Saha[uuuuuuuu], Orkun Saka[vvvvvvv], Chiara Saponaro[wwwwwwww], Georg Sator[xxxxxxxx], Martijn Schoenmakers[bb], Raffaello Seri[yyyyyyyy], Meet Shah[vvvv], Paul Sibille[yyyy], Christoph Siemroth[zzzzzzzz], Vladimir Skavysh[aaaaaaaaa,bbbbbbbbb], Ben Slater[ccccccccc], Wenting Song[ddddddddd], Stefan Staubli[sssssss], Tobias Steindl[eeeeeeeee], Nomwendé Steven Waongo[a], Paul Stott[fffffffff,ggggggggg], Stephenson Strobel[fffff], Roshini Sudhaharan[hhhhhhhhh], Pu Sun[iiiiiiiii], Scott D. Swain[jjjjjjjjj], Oleksandr Talavera[kkkkkkkkk], Hanz M. Tantiangco[lllllllll], Georgy Tarasenko[mmmmmmmmm], Boyd Tarlinton[nnnnnnnnn], Mariam Tarraf[l], Ken Teoh[ooooooooo], Rémi Thériault[ppppppppp,b], Bethan Thompson[qqqqqqqqq], Tonghui Tian[l], Wenjie Tian[a], Emmanuel Tolani[rrrrrrrrr,sssssssss], Nicolai Borgen[ttttttttt], Solveig Topstad Borgen[kkkkk], Javier Torralba[ooo], Carolina Velez-Ospina[uuuuuuuuu], Man Wai Mak[l], Lukas Wallrich[vvvvvvvv], Zeyang Wang[wwwwwwwww], Leah Ward[xxxxxxxxx], Matthew D. Webb[l], Duncan Webb[yyyyyyyyy], Bryan S. Weber[zzzzzzzzz,aaaaaaaaaa], Christoph Weber[bbbbbbbbbb], Wei-Chien Weng[cccccccccc], Christian Westheide[dddddddddd,eeeeeeeeee], Tom Wilkinson[xxxx], Kwong-Yu Wong[ffffffffff], Marcin Wroński[gggggggggg], Zhuangchen Wu[kkkkkkkkk], Qixia Wu[a], Victor Y. Wu[hhhhhhhhhh], Bohan Xiao[a], Feihong Xu[iiiiiiiiii], Cong Xu[jjjjjjjjjj,kkkkkkkkkk], Pranav Yadav[llllllllll], Yu Yang Chou[hhhhh], Luther Yap[mmmmmmmmmm], Myra Yazbeck[a,nnnnnnnnnn], Bo Yao[oooooooooo], Zuzanna Zagrodzka[pppppppppp], Tahreen Zahra[l], Mirela Zaneva[qqqqqqqqqq], Xiaomeng Zhang[rrrrrrrrrr], Ziwei Zhao[ssssssssss,tttttttttt], Han Zhong[uuuuuuuuuu], Aras Zirgulis[zz], Jiacheng Zou[vvvvvvvvv], Floris Zoutman[cccc], and Christelle Zozoungbo[wwwwwwwwwww]

Affiliations are included on p. 9.

Edited by Timothy D. Wilson, University of Virginia, Charlottesville, VA; received September 22, 2025; accepted March 16, 2026

Large Language Models (LLMs) such as ChatGPT are transforming how scientists conduct and validate research, offering promise as tools to improve scientific reproducibility. However, computational reproducibility and error detection remain expensive and labor-intensive. We experimentally test how collaboration between researchers and LLM assistants influences the reproduction of quantitative social science findings across different levels of AI autonomy. We randomly assigned 288 researchers to 103 teams working under three conditions: human-only, AI-assisted (using ChatGPT as a collaborative tool), or AI-led (ChatGPT operating with minimal human oversight). Teams reproduced published results from leading social science journals, detected coding errors, and proposed robustness checks. Human-only and AI-assisted teams achieved comparable reproduction rates (94% vs. 91%) and performed similarly on most outcomes, except human-only teams identified significantly more major coding errors. Both substantially outperformed AI-led teams, which achieved only a 37% reproduction rate, detected fewer errors across all categories, proposed weaker robustness checks, and required more time. This autonomous approach, however, likely represents only a lower bound of AI capabilities. Despite rapid model advances, expert human judgment currently remains indispensable for reliable empirical verification. While AI assistance did not degrade most outcomes, it provided no measurable advantages and was associated with reduced detection of major errors. However, the 37% autonomous reproduction rate indicates that AI could provide value in settings where scale or cost constraints preclude human review of papers, even though general-purpose LLMs offer no immediate advantages for human-supervised verification.

AI | reproducibility | large language models

Reproducibility is a cornerstone of robust quantitative empirical research, where complex methodologies and data handling techniques are common (1–8). Despite advancements in reproducibility protocols (9), concerns persist regarding the accuracy and reliability of published findings (10–17). Unclear reporting and methodological advances requiring expertise when evaluating quantitative studies contribute to the current reproducibility and replication crises in the behavioral and social sciences. At the same time, verifying computational reproducibility remains costly and labor-intensive (18). Even when journals require replication packages, reproducing results often involves navigating complex scripts, large datasets, and intricate empirical workflows. As empirical research becomes increasingly complex, scalable approaches to verification are needed to ensure that the reliability of published findings can be efficiently assessed.

This study investigates how artificial intelligence (AI) tools, such as Large Language Models (LLMs), could support researchers, data editors, and scientific journals in computationally reproducing research. We focus on three modes of AI and human interaction: human-only teams, human teams with AI assistance (the "AI-assisted" approach), and teams that provided only limited oversight while AI carried out reproducibility checks (the "AI-led" approach). The AI-led approach approximates a "protoagentic" system: an LLM tasked with reasoning through a reproducibility exercise with minimal human supervision. We use ChatGPT because it processes different file formats effectively for reproduction and is used most frequently by researchers (19).

This paper tests how effectively AI supports reproduction of scientific articles and works in complex cases where coding errors or methodological inconsistencies arise. We employ a randomized controlled trial design involving three treatment arms. We contribute to a large literature documenting the benefits and limitations of human–AI integration, as well as full automation (20). Evidence from human–AI decision-making suggests that performance ordering between AI alone, human alone, and human–AI teams is mixed and task-dependent, and that human–AI combinations often fail to outperform AI alone, sometimes performing worse due to miscalibrated trust and under- or overreliance on AI assistance (21–34). This is crucial for science because current methods for performing computational reproducibility and robustness checks are expensive, time consuming, and require advanced technical skills (18, 35). We also contribute to a growing body of literature documenting the potential pitfalls of integrating human and artificial intelligence, such as overreliance and expertise erosion (36, 37). This research also provides some comparative productivity measures in highly specialized intellectual tasks. This line of research mainly focuses on customer support agents and low-skill occupations, whereas we study high-skill scientific reproducibility tasks (29, 38).

Author contributions: A.B., D.V., A. Marcoci, J.P.A., D.M., B.B., R. Alexander, G.B., G.A.B., F.A., F.J.B.-B., L.-P.B., C. Cornejo, M.E., L. Fiala, J. Fitzgerald, R.F., J. Frese, G.G., A.G.-K., N.M., S. Merz, A. Musulan, A. Nordstrom, D.A.R., G.S., R. Seri, V.S., P. Sun, G.T., N.T.B., S.T.B., B.S.W., and L.Y. designed research; A.B., D.V., A. Marcoci, J.P.A., D.M., B.B., R. Alexander, L. Deer, T. Stafford, L.V., G.B., F.M., M. Abdelhady, Y.A., G.A.B., T. Aguirre, S. Aiyer, S. Akhtar, F.A., M.R.A., M. Altman, D.A., Z.A.L., J.A.d.L.T., D.R.A., I.A., A.-M.N., R. Ashong, T. Auer, F.J.B.-B., B.J.B., S.M.B., D.B., L.B., T.B., M.B., L.-P.B., C.G.B., C.B.A., C. Bolibaugh, C. Bonander, R.B., E.B., S.B., N.B., S.C.-S., A. Calef, J.S.C.A., G.A.C.A., S. Caulker, S. Cepenas, A. Chatton, Z.C., N.C.E., A.-B.C., F.J.C., J. Collins, N.C., C. Cornejo, J. Craveiro, J. Créchet, J. Cui, N.C.V., C. Czymara, C.D.B.J., H.D., L. Denoo, A.D., N.D., E. Djemai, E. Dujeancourt, U.D., T.D., Y.E., Y.E.F., I.E.F., K.E., A. Elminejad, M.E., A. Emirmahmutoglu, G.E.-F., E.E., J.F.D., J. Feld, A.F.R.J., G.F., V.F., L. Fiala, L. Fink, M.F., S.F., J. Fitzgerald, R.F., A.F.-C., L. Fréget, J. Frese, J.G., S.G., M.C.G., A. Gáspár, R.G., E.G., D. Geraldes, G.G.C., G.G., D. Goldschmitt, A.G.-K., A.G.d.V., I.G., A. Gugushvili, A.H.A.F., F.H., M. Hablicsek, J.H., J.D.H., O.H., M. Hassouneh, C.I.H., S.C.F.H., M. Hepplewhite, A.T.Y.H., S.H.-H., E.H., G.H., H.H., Z.G.I., P.J.S., N.J., J.J., E.J., H. Jebeli, Y.J., H. Junaid, R.K., S. Karim, E. Kelly, E. Kimel, S. Kingsuwankul, V.K., D.K., P.K., N.K., E. Kujansuu, C.F.K., S. Küster, B.L.-W., F.L., T.L., R.L., D.L., J.L., H.L., K.L., F.M.D., C.R.M., N.M., M. Mandas, C.M., J.M., D.M.F., X.M., R.M.W., D.M.-G., S. Meng, L. Meng, S. Merz, A.P.M., T.M., D.D.M., S. Mishra, B.W.M., M.M.M., M. Mohnen, L.-P.M., L. Muehlenbachs, G.M., A. Musulan, S. Muzzì, J.A.C.M., F.N., T.N., A. Niazi, A. Nordstrom, B.N., D.O., T.Ö., J.O., V.O.C., Ö.Ö., A.I.O., M.P., S.P., V.P., R.P., L.P., U.P., P.P., Q.Q., J.Q., D.A.R., J.R., N.R., S. Saha, O.S., C. Saponaro, G.S., M. Schoenmakers, R. Seri, M. Shah, P. Sibille, C. Siemroth, V.S., B.S., W.S., S. Staubli, T. Steindl, N.S.W., P. Stott, S. Strobel, R. Sudhaharan, P. Sun, S.D.S., O.T., H.M.T., G.T., B. Tarlinton, M.T., K.T., R.T., B. Thompson, T.T., W.T., M.T.R., E.T., N.T.B., S.T.B., J.T., C.V.-O., M.W.M., L. Wallrich, Z. Wang, L. Ward, M.D.W., D.W., B.S.W., C. Weber, W.-C.W., C. Westheide, T.W., K.-Y.W., M.W., Z. Wu, Q.W., V.Y.W., B.X., F.X., C.X., P.Y., Y.Y.C., L.Y., M.Y., B.Y., Z. Zagrodzka, T.Z., M.Z., X.Z., Z. Zhao, H.Z., A.Z., J.Z., F.Z., and C.Z. performed research; A.B., D.V., J.P.A., G.B., Z.A.L., I.A., A.-M.N., T. Auer, C.G.B., C. Bonander, R.B., S.B., A. Chatton, A.D., E. Dujeancourt, Y.E., J. Fitzgerald, O.H., A.T.Y.H., G.H., H.H., E. Kelly, V.K., N.K., R.L., J.L., K.L., S. Merz, S. Mishra, S. Muzzì, F.N., T.N., U.P., P. Sibille, S. Staubli, O.T., B. Tarlinton, M.T., M.T.R., C. Weber, Z. Wu, and V.Y.W. analyzed data; L. Deer, T. Stafford, G.B., F.M., S. Aiyer, S. Akhtar, F.A., M. Altman, I.A., A.-M.N., B.J.B., L.B., M.B., C.G.B., C. Bonander, R.B., S.B., A. Chatton, F.J.C., J. Collins, N.C., L. Denoo, E. Djemai, T.D., I.E.F., A. Elminejad, J.F.D., J. Feld, G.F., J. Fitzgerald, L. Fréget, J.G., S.G., A. Gáspár, E.G., G.G., D. Goldschmitt, A.G.-K., I.G., F.H., M. Hablicsek, J.D.H., S.C.F.H., A.T.Y.H., G.H., Z.G.I., J.J., E. Kelly, E. Kimel, S. Kingsuwankul, N.K., E. Kujansuu, S. Küster, B.L.-W., D.L., J.L., K.L., C.M., X.M., R.M.W., S. Merz, B.W.M., M. Mohnen, L.-P.M., L. Muehlenbachs, F.N., Ö.Ö., A.I.O., S.P., U.P., P.P., Q.Q., D.A.R., J.R., N.R., O.S., R. Seri, P. Sibille, S. Staubli, P. Sun, G.T., R.T., T.T., M.T.R., L. Wallrich, M.D.W., C. Weber, W.-C.W., M.W., V.Y.W., L.Y., B.Y., Z. Zagrodzka, M.Z., X.Z., H.Z., A.Z., and F.Z. edited and reviewed the paper; A.B., D.V., J.P.A., D.M., R. Alexander, L. Deer, T. Stafford, and L.V. organized replication games; and A.B., D.V., A. Marcoci, J.P.A., D.M., B.B., G.A.B., and L. Fiala wrote the paper.

Competing interest statement: The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada. A. Marcoci is a UK Research and Innovation 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.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: abrodeur@uottawa.ca.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2524747123/-/DCSupplemental.

Published May 28, 2026.

We focus on three groups of outcomes across the treatment arms: 1) computational reproducibility (success rate and time required), 2) error detection capabilities, and 3) proposing and implementing quality robustness checks. Understanding the impact of the treatment on these outcomes contributes to a broader understanding of AI, and offers insights into the optimal balance of human and AI involvement in research tasks.

1. Procedures

The first 10 coauthors organized seven AI replication games between February and November 2024, including a pilot in February. All remaining coauthors and a few of the organizers participated in one of those games. The participating coauthors were a mix of master and PhD students, postdoctoral fellows, professors, and researchers from nonacademic organizations with a doctoral degree. *SI Appendix*, Table S8 provides details on team composition. Randomization was carried out in two steps for each of the seven events. In step one, participants 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 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: human-only, AI-assisted, or AI-led.

Each team was assigned a study from leading social science journals (i.e., economics, political science, or behavioral science/psychology). Each event included two studies with known coding errors (one in Stata and one in R) that had been identified by the lead authors in a prior study but were not publicly disclosed at the time of the AI replication game. Detailed information about the papers used that contained coding errors can be found in *SI Appendix*, Tables S1 and S2. Descriptions of the coding errors identified prior to each replication game can be found in *SI Appendix*, Table S3 through *SI Appendix*, Table S4. Coding errors occurred when preparing data for analysis (variable definitions, incorrect merging of datasets, differing sample restrictions, not cleaning variables, missing variables) as well as when carrying out the analysis (discrepancies between code and what is written in the article). Examples of the latter include inconsistently specified SEs and control variables. Teams and local organizers had no information about the study they would be reproducing until the start of the event. Twelve studies were used in total, with a few reused for multiple events.

Relevant resources were given to the teams at 09:00 local time on the day of the event. We shared with them: the journal article and online appendix as PDFs, the original authors' replication package, and screenshots of the exhibit to reproduce from the article (*SI Appendix*). Screenshots were introduced after the pilot event to assist AI-led teams, as the AI might be better able to process tables and figures as images rather than when embedded in PDF files. Teams had seven hours to complete three tasks: i) computationally reproduce a few predetermined results, ii) detect coding errors, and iii) suggest and implement up to two robustness checks. The three tasks were independent from each other (e.g., teams did not need to fix coding errors to computationally reproduce results). However, teams were instructed to begin with reproducing the results before proceeding to specifically search for coding errors and propose robustness checks. Teams could leave before the end of the event if they believed they had completed their tasks as feasibly as possible. Upon completion, teams were asked to email the lead authors a (templated) time log documenting whether they completed computational reproducibility, with a list of all coding errors uncovered, and two robustness checks. All AI-assisted and AI-led teams used ChatGPT during the event and had to provide their AI conversation history (i.e., a transcript of all prompts and responses exchanged with ChatGPT).

Participants were offered coauthorship on this paper, independent of their team's performance or success in reproducing results. No monetary compensation or performance-based incentives were provided. While this may have led to reduced effort for some teams, it also reduced incentives for strategic behavior or protocol violations, particularly for AI-led teams who were asked not to directly examine the article, code, or data.

Access to a paid subscription of ChatGPT (powered initially by GPT-4 and subsequently by other models) was provided to all members in the AI-assisted and AI-led teams. While ChatGPT had six different versions available between February 14th 2024 (training for our pilot) and our final event on November 22nd 2024, researchers had access to the main flagship models (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. Additional information on the

Significance

Verifying results of published social sciences research is essential but expensive, costing hundreds of dollars per study. With AI tools like ChatGPT becoming widespread, we tested whether they could help scientists check if research findings can be reproduced. We assigned 288 researchers to 103 teams working with no AI, with AI as an assistant, or AI leading the work with minimal human input. Human teams and AI-assisted teams performed similarly on most tasks, but humans caught more critical errors. AI working autonomously achieved a 37% reproduction rate, making it potentially useful for automated screening when human review is cost-prohibitive. These results nonetheless show that human expertise remains essential for reliable scientific validation.

different versions and use by teams at events are included in *SI Appendix*, ChatGPT Models.

AI-assisted and AI-led teams took part in a mandatory one-hour training on the usage of ChatGPT. Participants viewed the training live or later via recording. The AI training was optional for human-only teams. The training had nine components which we outline here but describe further in *SI Appendix*, AI Training: 1) Introduction, Overview of ChatGPT and Access; 2) Interaction with ChatGPT; 3) Sharing Chats with I4R; 4) Coding Assistance; 5) Uploading Files and Images; 6) Conducting Data Analysis Using ChatGPT; 7) ChatGPT API; 8) Customizing ChatGPT; and 9) Explanation of Differences Among ChatGPT Models. AI-led and AI-assisted teams constructed their own prompts but were given examples and best-practice guidance in the training session. Using textual analysis on all prompts, we show limited overlap in prompt wording across AI-led and AI-assisted teams (*SI Appendix*).

The human-only teams were not allowed to use ChatGPT or any other AI tool. The AI-assisted teams were allowed to use ChatGPT without limitation (but no other AI tool). AI-led teams had to perform the tasks only using the guidance of ChatGPT. They were not allowed to read the article or look at the data and code but could ask ChatGPT to summarize the article. They had to upload the article to ChatGPT along with an image of the table(s)/figure(s) to be reproduced, the replication code, and the data files where feasible. They were asked to first 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 the one provided by ChatGPT, though the teams could adjust file paths and their environment without the assistance of ChatGPT. During the pregames AI training, participants were shown examples of how to upload the article and replication files to ChatGPT and how to use the Python interpreter module. We relied on the integrity of the AI-led teams to not look at the studies, code, or files. That is, we asked them to pass everything through ChatGPT; we did not give specific guidance on how teams should operate. Teammates could work independently or jointly throughout the event.

In summary, we have 103 teams: 33 human-only teams (92 researchers), 35 AI-assisted teams (93 researchers), and 35 AI-led teams (103 researchers). *SI Appendix*, Table S8 shows the treatment arms are balanced across observables.

1.1. Three Tasks. Participants had three objective tasks with measurable outcomes. First, teams were asked to computationally reproduce a few selected results in the study assigned to them. The numerical results were selected by the lead author, AB, based on their relative importance to the main claims of the article. Computational reproducibility involves using the same data as the original authors and running their code. In the templated log, teams recorded the time taken to computationally reproduce the numerical result. Notably, AB, JA, and DM were able to computationally reproduce the selected results before the event, requiring only minimal adjustments (e.g., updating file paths). We have two different dependent variables for computational reproducibility: one outcome as a binary variable (completed computational reproducibility vs. did not complete), and one that is time (in minutes) from the start of the event to when teams completed a computational reproduction. A computational reproduction is defined as the successful execution of the original authors' codes and the production of numerical results in line with those in the article.

Second, we compare how effective different team types were in finding coding errors or data irregularities. For simplicity, we refer to these as "errors." We categorize errors as major or minor based on whether they 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 is considered a major error, as it could have an impact on the estimation results. In contrast, minor coding errors are typically easily fixed by the reproducers and do not impact the validity of the claims made by the original authors. In a set of exploratory analyses, we also categorize coding errors along three dimensions: i) whether the error occurs in preparing the data and analysis, ii) whether the error is related to the regression analysis, and iii) whether it is a transcription error (e.g., a mismatch between the coefficient reported in the article and the coefficient produced by the code, such as −0.034 vs. 0.034). We also investigate the extent of false error detection and the share of errors not uncovered by the treatment arm.

Third, we asked each team to propose and perform two robustness checks. A robustness check is defined in our study as an additional statistical computation. We instructed that these robustness checks should not repeat ones already mentioned in the study or its *SI Appendix*, that they should be feasible, and that heterogeneity analysis (e.g., comparing female and male respondents) was not considered a robustness check.

Defining what makes a robustness check "good" or "bad" is not straightforward. We define four binary criteria for evaluating the quality of robustness checks: i) clarity of purpose and execution; ii) feasibility; iii) novelty (i.e., not previously done by the original authors); and iv) relevance to the validity of the empirical strategy. Items i) through iii) are basic necessary conditions. Item iv) requires that the purpose of the robustness test is to provide evidence regarding the credibility of the empirical strategy (39–41). All four criteria must be met for a robustness check to be considered "good." Additionally, running corrected code in an attempt to correct major errors in the original paper is coded as a "good" robustness check, regardless of whether it complies with the previous criteria.

We measure differences by team type in proposing and implementing robustness tests using four measures. The first two are whether teams proposed one or two "good" robustness checks. The third and fourth dependent variables are whether the participants report to have implemented one or two of those "good" robustness checks, respectively.

2. Results

Our analyses were preregistered after the pilot event in Toronto. We list deviations from our preregistration in *SI Appendix* and note throughout whether the analysis is exploratory.

2.1. Computational Reproducibility. Our main finding is that computational reproducibility rates varied substantially across the groups. Most human-only (94%; 31/33) and AI-assisted (91%; 32/35) teams could computationally reproduce the results, while only 37% (13/35) of AI-led teams were able to do so (Table 1). Table 2 shows the ordinary least squares (OLS) estimates of our main regression model (see *SI Appendix*, Table S10 for logit and Poisson regressions and *SI Appendix*, Table S12 for coefficient estimates concerning the control variables). We find that human-only teams are about 59 percentage points more likely than

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	82.0 (39.8)	93.3 (85.4)	179.7 (68.4)	-11.3 [0.505]	-97.7 [<0.001]	-86.4 [0.002]
Number of minor errors	1.000 (1.658)	1.400 (2.488)	0.686 (1.605)	-0.400 [0.441]	0.314 [0.430]	0.714 [0.158]
Minutes to first minor error	141.6 (97.0)	139.9 (83.1)	157.6 (85.6)	1.7 [0.960]	-16.0 [0.691]	-17.7 [0.622]
Number of major errors	1.697 (2.568)	0.743 (1.120)	0.229 (0.547)	0.954 [0.049]	1.468 [0.002]	0.514 [0.017]
Minutes to first major error	110.5 (69.5)	130.3 (86.9)	152.8 (94.3)	-19.8 [0.487]	-42.3 [0.261]	-22.5 [0.606]
At least one good robustness check	1.000 (0.000)	1.000 (0.000)	0.829 (0.382)	0.000 [NA]	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: Columns 2–4 present means and SEs in parentheses for individual groups (Human-only, AI-Assisted, and AI-Led); columns 5–7 present differences in means and *P*-values in brackets for group comparisons (Human-Only vs. AI-Assisted, Human-Only vs. AI-Led, and AI-Assisted vs. AI-Led).

AI-led teams to successfully computationally reproduce the results ($P < 0.001$). In contrast, there is no statistically significant difference between human-only and AI-assisted teams ($P = 0.771$).

We next investigate how the distribution of time-to-computational reproduction varies across groups. Fig. 1 plots complementary Kaplan–Meier curves showing, by treatment arm, how long teams took to reproduce their paper by the end of the event. The proportion of teams that reproduce their paper does not reach 100% after seven hours in any treatment arm because all treatment arms contain some teams who could not reproduce their paper. This is especially noticeable for AI-led teams. We find that human-only and AI-assisted teams are significantly faster than AI-led teams (Table 1). There is no statistically significant difference between human and AI-assisted teams.

In an exploratory analysis, we investigate whether AI-assisted and AI-led teams improved over time. In our setting, improvements could be due to new ChatGPT versions and increased researchers' skills over time. In *SI Appendix*, Fig. S2, we show the difference in computational reproducibility rates between the treatment groups by event. Visually, AI-led teams did not improve over time when compared to human-only teams during the first five events in 2024. We observe that the reproducibility

Table 2. Causal relationship between treatment groups and reproducibility 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	0.313	-1.022***	-0.009	-0.014	-0.032	-0.009
	(0.063)	(0.387)	(0.362)	(0.027)	(0.103)	(0.061)	(0.113)
	[-0.144; 0.107]	[-0.458; 1.083]	[-1.743; -0.300]	[-0.063; 0.046]	[-0.220; 0.191]	[-0.155; 0.090]	[-0.233; 0.216]
AI−led	-0.593***	-0.331	-1.344***	-0.167**	-0.250**	-0.323***	-0.290**
	(0.090)	(0.350)	(0.342)	(0.068)	(0.107)	(0.098)	(0.126)
	[-0.773; -0.413]	[-1.029; 0.366]	[-2.024; -0.664]	[-0.302; -0.031]	[-0.463; -0.037]	[-0.518; -0.127]	[-0.540; -0.040]
Controls	✓	✓	✓	✓	✓	✓	✓
Mean of dep. var	0.738	1.029	0.874	0.942	0.786	0.816	0.680
P-val (AI-assisted = AI-led)	0.000	0.115	0.251	0.021	0.032	0.003	0.017
Obs.	103	103	103	103	103	103	103

Note: SEs in parentheses; CIs in brackets. Human-only group omitted.

Controls: number of teammates; game-by-software fixed effects; maximum and minimum position skill fixed effects; attendance fixed effects.

$^{*}P < 0.10$, $^{**}P < 0.05$, and $^{***}P < 0.01$.

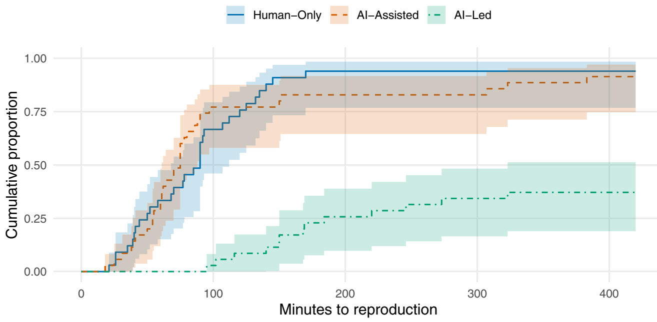

Fig. 1. Complementary Kaplan–Meier curves, showing the proportion of teams who computationally reproduced the paper by time *t* along with curve bands (95% CIs).

rate gap between human-only and AI-led teams was over 50 percentage points for most events in 2024. Of note, this gap had slightly narrowed by the final event of 2024.

2.2. Coding Errors or Data Irregularities. We have two primary dependent variables concerning coding error detection: counts of major and minor errors detected. We find that human-only teams identified on average 1.00 minor and 1.70 major errors, compared with 1.40 minor and 0.74 major errors for AI-assisted teams and 0.69 minor and 0.23 major errors for AI-led teams, respectively (Table 1). Table 2 provides OLS estimates indicating that, compared to AI-assisted and AI-led teams, human-only teams uncovered more major errors ($P = 0.006$ and $P < 0.001$, respectively). The difference in the number of minor coding errors detected is, however, not significant ($P = 0.421$ and $P = 0.347$, respectively). We further find that AI-assisted teams uncovered more minor errors than AI-led teams, but the estimate is not significant at any conventional level ($P = 0.115$). *SI Appendix* provides examples of errors and a discussion.

Fig. 2 plots complementary Kaplan–Meier curves showing how long teams took to find a first minor error (*Top* panel) and a first major error (*Bottom* panel). We find that the speed at which AI-assisted teams uncover a first (minor or major) error is not statistically significantly different from that of human-only teams and that AI-led teams are statistically significantly slower than human-only teams at uncovering a first major error.

Our findings suggest that human-only 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 detect data irregularities.

In exploratory analyses, we explore whether AI-led and AI-assisted teams are better at uncovering different types of errors, distinguishing between coding mistakes that require substantive understanding of the paper and those that do not. Table 3 provides OLS estimates indicating that, compared to AI-led teams, human-only teams uncovered more errors that occur in preparing the data and analysis (although not significantly different, $P = 0.165$), more errors related to the regression analysis ($P = 0.007$) and more transcription errors ($P = 0.034$). Human-only teams uncover more errors in these three categories than AI-assisted teams, but only one of the point estimates is statistically significant at the 10% level ($P = 0.993$, $P = 0.072$, and $P = 0.318$).

Our qualitative analysis (Section 2.5) suggests that some AI-led teams experienced prompt fatigue and hallucinated reasoning paths. This qualitative evidence motivates our exploratory analysis of whether AI-assisted and AI-led teams are more likely to produce false error detection. We find no evidence that this is the case ($P = 0.248$, $P = 0.642$), suggesting that hallucinations occur at other stages of the reproduction pipeline. This previous result may mask the fact that many AI-led teams did not detect any errors. We thus investigate the proportion of errors that remain undetected by each team. We find that AI-led teams detected a significantly smaller proportion of known errors than human-only and AI-assisted teams ($P < 0.001$, $P = 0.016$). These results suggest that AI-led teams' primary limitation lies in error discovery rather than erroneous overdetection.

We also provide noncausal evidence in exploratory analyses in *SI Appendix*, Table S14 that AI-assisted teams with more AI experience uncovered coding errors faster, although these estimates are only statistically significant at the 10% level ($P = 0.067$ and $P = 0.070$). Extending this analysis, *SI Appendix*, Table S15 compares human-only teams with AI-assisted teams with high vs. low/medium AI experience. These comparisons should be interpreted with caution, as the number of AI-assisted teams in each subgroup is small, resulting in limited statistical power. Nonetheless, the point estimates are consistent with the hypothesis that AI experience improves the effectiveness of AI-assisted teams. AI-assisted teams with high AI experience appear to uncover coding errors faster than human-only teams and detect more minor errors on average. The magnitudes of these differences are sizable, but the estimates are imprecise and not statistically significant at conventional levels. These findings are consistent with the behavioral evidence presented in Section 3.4, which examines how AI-assisted teams used ChatGPT.

We also find that Stata teams uncovered significantly more major errors ($P < 0.001$), with the human-only groups using Stata finding significantly more major errors than all other groups (*SI Appendix*, Table S13).

In an exploratory analysis, we investigate if the performance of AI-led teams in detecting errors improved over time. *SI Appendix*, Figs. S4 and S6 suggest no improvement of AI-led teams relative to human-only teams over the year 2024.

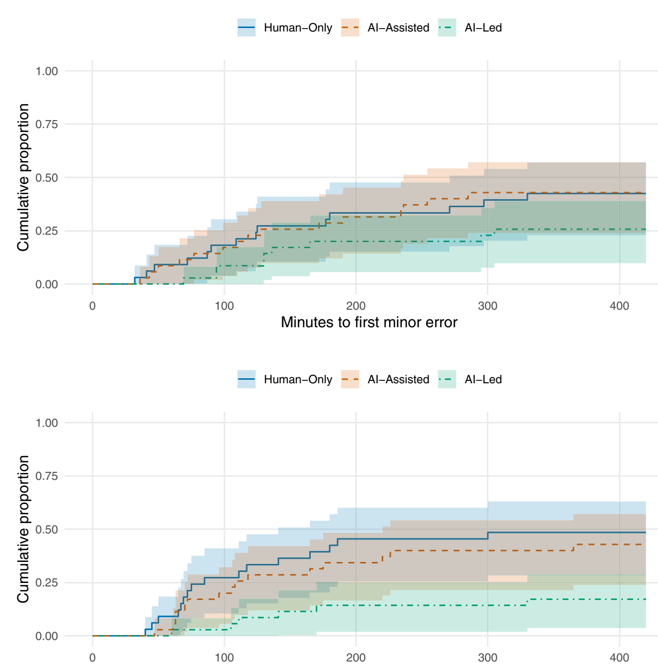

Fig. 2. Complementary Kaplan–Meier curves, showing the proportion of teams who found their first coding error by time *t* along with curve bands (95% CIs).

Table 3. Causal relationship between treatment groups and error types

	(1) Preregression errors	(2) Regression errors	(3) Transcription/postregression errors	(4) False error detection	(5) Share of known errors not found
AI-assisted	-0.002	-0.604*	-0.343	-0.359	0.042
	(0.267)	(0.331)	(0.341)	(0.309)	(0.048)
	[-0.534; 0.530]	[-1.263; 0.055]	[-1.023; 0.337]	[-0.974; 0.256]	[-0.053; 0.137]
AI-led	-0.407	-0.886***	-0.652**	0.221	0.151***
	(0.290)	(0.321)	(0.301)	(0.473)	(0.043)
	[-0.986; 0.171]	[-1.524; -0.248]	[-1.251; -0.052]	[-0.721; 1.162]	[0.065; 0.236]
Controls	✓	✓	✓	✓	✓
Mean of dep. var	0.786	0.660	0.583	0.680	0.846
P-val (AI-assisted = AI-led)	0.140	0.228	0.274	0.146	0.016
Obs.	103	103	103	103	103

Note: SEs in parentheses; CIs in brackets. Human-only group omitted.
Controls: number of teammates; game-by-software fixed effects; maximum and minimum position skill fixed effects; attendance fixed effects.
*$P < 0.10$, **$P < 0.05$, and ***$P < 0.01$.

2.3. Proposed Robustness Checks. We find a clear, consistent performance hierarchy across both conditions: Human-only and AI-assisted teams outperform AI-led teams. We find that all human-only (33/33) and AI-assisted (35/35) teams proposed at least one good robustness check, whereas only 83% (29/35) of AI-led teams did so. Table 2 provides OLS estimates and show that the difference between AI-led groups and the other two groups is statistically significant ($P = 0.017$ and $P = 0.021$, respectively). We find that 29 of 33 human-only and 30 of 35 AI-assisted teams suggested two good checks, compared with just 22 of 35 AI-led teams (Table 2, $P = 0.022$ and $P = 0.032$).

Looking at whether teams report to have implemented those checks, AI-led teams were almost 32 percentage points less likely than the other two groups to report having conducted a robustness check that was classified as "good" ($P = 0.002$ and $P = 0.003$), and six AI-led teams supplied no robustness checks evaluated as "good" at all. These six teams' checks were judged as "bad" mostly because of a lack of clarity and duplicating analyses already run by the original authors.

Our results indicate that AI-led teams, while able to produce robustness checks with some level of quality, faced more challenges in aligning with the criteria. These difficulties may stem from omission of relevant information when describing the task to the AI or from limited ability of the AI to interpret the empirical strategy and to assess the feasibility of the checks.

2.4. Additional Analyses for AI-Assisted Teams. *SI Appendix*, Table S16 presents an exploratory correlational analysis examining the relationship between AI usage (measured by total prompts) and performance in AI-assisted teams. See *SI Appendix*, Fig. S7 for descriptive statistics on AI usage for AI-assisted teams. For this analysis, we divided teams into lower and higher AI-usage groups using a median split based on the total number of prompts they employed.

The findings indicate that AI-assisted teams with lower AI usage were less likely to achieve computational reproduction of the original results and uncovered less major and minor coding errors. Of note, our sample is small and none of the differences are statistically significant at the 5% level. To further explore potential mechanisms behind this heterogeneity, *SI Appendix*, Table S14 also reports differences in prompting behavior by AI experience. We find that AI-assisted teams with lower AI experience tend to interact with ChatGPT more frequently, as measured by the number of prompts, but the difference is not statistically significant due to the small sample size ($P = 0.307$). This pattern is consistent with the idea that less experienced teams rely more heavily on iterative prompting, which may contribute to longer task completion times. These results relate to a literature studying overreliance on AI support (20, 36, 37, 42–44).

2.5. Focus Groups. In additional exploratory analysis, between 18 April and 30 April 2025, we conducted six one-hour focus groups ($n = 25$) involving AI-led ($n = 8$), AI-assisted ($n = 11$), and human-only ($n = 6$) participants. The participants were aware of the headline quantitative results. While this creates risk of confirmation bias and demand characteristics, we addressed this by emphasizing process, task allocation, and failure points rather than outcomes in the discussion guide, and by treating focus group material as explanatory and triangulatory evidence rather than independent support for treatment differences. Accordingly, we use qualitative themes to illuminate mechanisms behind observed patterns, and we flag any tensions between participant claims and the experimental results. Consistent with this stance, where participant views exceeded what the quantitative results support, we report the discrepancy rather than treat it as confirmation.

Thematic analysis revealed the following patterns: AI-assisted participants reported that AI assistance sped things up, while AI-led participants reported the opposite. Human-only participants believed they were the most effective at detecting major errors, with AI-led participants trailing. AI-assisted teams strategically outsourced microtasks, for example boilerplate code and file location, while retaining conceptual control, whereas AI-led teams were required to cede entire analytic stages to ChatGPT and struggled when automation failed.

Data illuminated the practical consequences of these differences. Initial optimism about LLMs quickly gave way to prompt fatigue by participants, model's overconfidence, and mounting frustration, especially among AI-led participants who faced hallucinated paths, truncated context windows, and prolonged debugging loops. AI-assisted teams report that human expertise remained necessary for detecting subtle errors and for arbitrating disagreements between AI output and reality. Nevertheless, when used judiciously, LLMs accelerated routine work, suggested robustness checks, and broadened analytical ambition for less experienced coders. Therefore, our focus group findings imply that effective LLM prompting is becoming a specialized research skill and that near term gains will come from augmenting, not replacing, human judgment. Additional details on methodology and results from the focus group analysis are provided in *SI Appendix*.

3. Discussion

Computational reproducibility, error detection, and robustness checks are essential components of empirical research validation, but are resource-intensive tasks. Ensuring that research can be reproduced is financially demanding. Recent research suggests that, across 10 top economics journals, the average expense of reproducing a single study is about USD $365 (18). For the American Economic Association, data-editor activities cost approximately USD $750 per article (9). Against this backdrop, our comparative analysis of human-only, AI-assisted, and AI-led teams sheds light on how AI may be integrated into the costly reproducibility pipeline, potentially accelerating some stages of the process and reshaping how replication labor is allocated.

A key finding of our study is that AI-led teams were able to successfully computationally reproduce approximately 37% of results. This result is nontrivial and suggests that a first automated pass at computational reproducibility was already within reach for a meaningful subset of empirical work in 2024.

At the same time, our results temper expectations of immediate, widespread AI autonomy in reproducibility. While recent advances in large language models have expanded the scope for AI integration in research workflows (45, 46), AI-led and AI-assisted teams do not yet outperform human-only teams on average. Moreover, current AI deployments introduce additional costs, such as paid model subscriptions, without consistently delivering higher success rates. As a result, fully autonomous AI reproduction does not yet offer clear cost savings relative to experienced human researchers.

However, our study likely represents only a lower bound of the capabilities of a more fully developed autonomous AI replication system. In practice, an AI system could deploy more sophisticated prompting strategies, exploit parallel experimentation, and possibly be supervised by trained research assistants or undergraduate students rather than senior researchers, reducing labor costs while maintaining acceptable levels of oversight. Our findings thus imply that future iterations of AI-led reproducibility systems may achieve higher success rates without proportional increases in human effort.

This perspective reframes AI not as a replacement for human expertise, but as a tool for redistributing effort across stages of the reproducibility pipeline. AI systems may handle routine debugging, error detection, and preliminary robustness checks (47–49), while human researchers focus on interpretation, judgment, and more complex failures. Under this model, even partial automation can generate meaningful cost savings and efficiency gains at scale.

3.1. Summary of Findings. AI-led teams 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 capacity of AI in 2024 to autonomously guide researchers through complex quantitative analyses. Similarly, in error detection, AI-led teams documented significantly fewer major and minor errors than either AI-assisted or human-only teams. These findings underscore the importance of still integrating human expertise. As LLMs continue to evolve, sustained benchmarking against humans will be crucial to ensure that future AI-led efforts close and potentially surpass the existing performance gap.

3.2. Limitations. One limitation is our sole focus on OpenAI's ChatGPT, 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 conducted depending on the field of science. In addition, participant incentives and attribution dynamics may have encouraged some teams to minimize time or effort, potentially increasing overreliance on AI tools. Finally, our analysis is based on a small, nonrandom set of studies spanning a limited range of social science methodologies and replication difficulty levels; although we provide detailed proxies for task complexity and error types, this sample composition constrains the extent to which our findings on AI assistance generalize across papers of different difficulty and across other scientific fields (*SI Appendix*, Table S5).

We note that participant behavior may have been influenced by observation and professional identity, generating a Hawthorne-type effect. Researchers with strong coding skills or a personal stake in reproducibility may have exerted greater effort in human-only teams, while responsibility may have been partially shifted to the AI in AI-assisted or AI-led settings. While this could bias relative performance comparisons, it may also reflect real-world incentive and attribution dynamics that shape how AI tools are adopted in research practice.

3.3. 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 late 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 human-only teams in overall outcomes but outperformed AI-led teams on most of our outcomes.

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, highlighting the limits of current AI in fully autonomous reproduction.

3.4. Outlook. Advancements in models and further optimization of AI for reproduction may soon address the limitations we reported. Future advancements in models optimized through reinforcement learning to solve reasoning problems using 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.

Future research should consider the potential for training models specifically in 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 (e.g., with native support for R and Stata) 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.

Future research should also focus on analyzing which prompting strategies leads to successful reproductions and which paths lead to failure, insights that could inform the development of AI systems better tailored for social science research. We make the chat transcripts publicly available and conduct an exploratory analysis of ChatGPT transcripts in *SI Appendix*.

4. Materials and Methods

Participants in the AI replication games experiments coauthor this study. The University of Ottawa Office of Research. Our preanalysis plan was preregistered on the Open Science Framework (OSF) on May 2nd, 2024, after our pilot event at the University of Toronto (https://osf.io/sz2g8/). AI-assisted and AI-led teams took part in a mandatory one-hour ChatGPT training, while the same training was optional for human-only teams; slides and recordings are available on OSF. A version-tagged copy of the code and data is permanently archived at https://github.com/I4Replication/AI-Games, and we make our AI training materials and recording, data and code, preanalysis plan, and template form available at https://osf.io/sz2g8/ with no restrictions on sharing or reuse.

5. Research Ethics Boards

Participants in the AI replication games experiments coauthor this study. The University of Ottawa Office of Research Ethics and Integrity reviewed and approved our AI games (H-09-25-12041). The King's College London Research Ethics Office reviewed and approved our focus groups (MRA-24/25-48393). All participants provided informed consent.

Data, Materials, and Software Availability. We make our i) AI training materials and recording, ii) data and code, iii) preanalysis plan and iv) template form available here: https://github.com/I4Replication/AI-Games (50). We declare no restrictions on sharing or reuse.

ACKNOWLEDGMENTS. We would like to thank Gabriel Zimmerman for research assistance. This research and AI replication games were funded by Coefficient Giving 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. S. Akhtar acknowledges funding DP200102935 awarded by the Australian Research Council Grant. R. Seri acknowledges funding from project "Dipartimento di Eccellenza 2023-2027" awarded by the Ministero dell'Università e della Ricerca.

Author affiliations: [a]Department of Economics, Faculty of Social Sciences, University of Ottawa, Ottawa, ON K1N 6N5, Canada; [b]Institute for Replication, Ottawa, ON K1N 6N5, Canada; [c]Institute for Technology and Humanity, University of Cambridge, Cambridge CB2 1SB, United Kingdom; [d]Department of Statistical Sciences, Faculty of Information, University of Toronto, Toronto, ON M5S 3G6, Canada; [e]Department of Management and Marketing, University of Melbourne, Melbourne, Carlton, VIC 3010, Australia; [f]School of Psychology, University of Sheffield, Sheffield S1 4DP, United Kingdom; [g]Research on Research Institute, London WC1E 6JA, United Kingdom; [h]Department of Economics, Cornell University, Ithaca, NY 14853; [i]Climate and Development Policy Division, Leibniz Institute for Economic Research - Leibniz Institute for Economic Research, Essen, NRW 45128, Germany; [j]Robert C. Vackar College of Business and Entrepreneurship, University of Texas Rio Grande Valley, Edinburg, TX 78539; [k]Norwich Business School, Accounting and Quantitative Methods, University of East Anglia, Norwich NR4 7TJ, United Kingdom; [l]Department of Economics, Carleton University, Ottawa, ON K1S 5B6, Canada; [m]Statistics Canada, Ottawa, ON K1A 0T6, Canada; [n]GovAI, N1 9JY, London, United Kingdom; [o]Department of Experimental Psychology, University of Oxford, Oxfordshire OX1 3EL, United Kingdom; [p]Finance Discipline, University of Sydney Business School, The University of Sydney, Sydney, NSW 2006, Australia; [q]Department of Actuarial Studies and Business Analytics, Macquarie Business School, Macquarie University, Sydney, NSW 2019, Australia; [r]Department of Socioeconomics, Vienna University of Economics and Business, Vienna 1020, Austria; [s]Center for Research on Equitable and Open Scholarship, Massachusetts Institute of Technology, Cabmridge, MA 02139; [t]Technical University of Munich School of Management, Technical University of Munich, Munich 80333, Germany; [u]Centre for Business Research, University of Cambridge, Cambridge CB2 1QA, United Kingdom; [v]Independent researcher, Shiraz, Iran; [w]Facultad de Economía, Universidad del Rosario, Bogotá 111711, Colombia; [x]School of Economics, Universidad del Rosario, Bogotá 111711, Colombia; [y]International Center for Higher Education Research and Faculty of Economics, University of Kassel, Kassel, Hessen 34125, Germany; [z]Department of Economics, University of Ghana, Legon, Accra, Ghana; [aa]Department of Economics, University of Basel, Basel 4052, Switzerland; [bb]Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg 5037 AB, The Netherlands; [cc]Department of Sport, Tourism and Hospitality Management, Temple University, Philadelphia, PA 19312; [dd]Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom; [ee]Center for Economic Policy Research, Coldbath Square, London EC1R 5HL, United Kingdom; [ff]Department of Economics, University of Zurich, Zurich CH-8001, Switzerland; [gg]Department of Behavioural Sciences and Learning, Linköping University, Linköping 58183, Sweden; [hh]Centre for Experimental Social Sciences, Nuffield College, University of Oxford, Oxford OX1 1NF, United Kingdom; [ii]Department of Economics, University of Southampton, Southampton SO17 1BJ, United Kingdom; [jj]Department of Finance, Norwegian School of Economics, Bergen 5045, Norway; [kk]Faculty of Psychology, Atma Jaya Catholic University of Indonesia, Jakarta 12930, Indonesia; [ll]Department of Education, University of York, York YO10 5DD, United Kingdom; [mm]Karlstad Business School, Karlstad University, Karlstad SE-651 88, Sweden; [nn]Center for Societal Risk Research, Karlstad University, Karlstad 65188, Sweden; [oo]Faculty of Biology Medicine and Health, The University of Manchester, Manchester M13 9NT, United Kingdom; [pp]School of Business and Economics, Maastricht University, Maastricht 6211 LM, The Netherlands; [qq]Department of Political Science, University of Chicago, Chicago, IL 60637; [rr]Centre for Environmental Sciences, Hasselt University, Hasselt 3500, Belgium; [ss]International Center for Higher Education Research, University of Kassel, Kassel, Hessen 34125, Germany; [tt]Meta-Research Innovation Center at Stanford, Stanford University, Stanford, CA 94305; [uu]Department of Economics, Nazarbayev University, Astana 010000, Kazaakhstan; [vv]Facultad de Economía, Universidad de los Andes, Bogotá 111711, Colombia; [ww]University College London School of Management, University College London, London E14 5AA, United Kingdom; [xx]Department of Economics, Universidad de Los Andes, Bogotá 111711, Colombia; [yy]United Methodist University Sierra Leone, Freetown, Sierra Leone; [zz]Department of Economics, International School of Management University of Management and Economics, Vilnius LT-01103, Lithuania; [aaa]Département de Médecine Sociale et Préventive, Université Laval, Québec, QC G1V0A6, Canada; [bbb]Département de Médecine Sociale et Préventive, Université de Montréal, Montréal, QC H3N1X9, Canada; [ccc]Department of Management, Marketing and Information Systems, Hong Kong Baptist University, Kowloon Tong, Hong Kong Special Administrative Regions of China; [ddd]Department of Economics, College of Management Sciences, Michael Okpara University of Agriculture, Umudike, Abia State 440109, Nigeria; [eee]Business School, University of Technology Sydney, Sydney, NSW 2007, Australia; [fff]Department of Economics, Wilfrid Laurier University, Waterloo, ON N2L 3C5, Canada; [ggg]Department of Medical Statistics, The London School of Hygiene & Tropical Medicine, London WC1E 7HT, United Kingdom; [hhh]Department of Computer Science, University College London, London WC1E 6BT, United Kingdom; [iii]Department of Economics, Faculty of Social Sciences, University of Ottawa, Ottawa ON, Canada; [jjj]Business School, Beijing Normal University, Beijing 100875, China; [kkk]Beijing Normal University, Belt and Road School, Zhuhai, Guangdong 519085, China; [lll]School of Earth and Planetary Science, National Institute of Science Education and Research, Odisha 752050, India; [mmm]Migration & Migrants, Netherlands Interdisciplinary Demographic Institute, The Hague NL-2511 CV, the Netherlands; [nnn]Department of Economics, Universidad del Rosario, Bogotá 111711, Colombia; [ooo]Department of Marketing, Tilburg University, Tilburg 5000 LE, The Netherlands; [ppp]Department of Strategy & Entrepreneurship, Tilburg University, Tilburg 5000 LE, The Netherlands; [qqq]Department of Economics, Binghamton University, Binghamton, NY 13902; [rrr]Economics, Université Paris Dauphine-Paris Sciences et Lettres, Paris CEDEX 16 75775, France; [sss]Swedish Institute for Social Research, Stockholm University, Stockholm 103 91, Sweden; [ttt]Department of Economics and Statistics, Linnaeus University, Växjö 35195, Sweden; [uuu]Department of Marketing, Vienna University of Economics and Business Vienna, Vienna 1020, Austria; [vvv]Financial Stability Department, Bank of Canada, Ottawa, ON K1A 0G9, Canada; [www]Onsi Sawiris School of Business, The American University in Cairo, New Cairo 11835, Egypt; [xxx]Department of Accounting and Control, Hautes études commerciales Lausanne, Lausanne 1015, Switzerland; [yyy]Institute for Accounting, Controlling and Auditing,, University of St. Gallen, St. Gallen 9000, Switzerland; [zzz]Monash University, Melbourne, VIC 3168, Australia; [aaaa]School of Psychology and Vision Science, University of Leicester, Leicester LE1 7RH, United Kingdom; [bbbb]Psychology, University of Birmingham, Birmingham B15 2TT, United Kingdom; [cccc]Department of Business and Management Science, Norwegian School of Economics, Bergen 5045, Norway; [dddd]Konjunkturforschungsstelle Swiss Economic Institute, ETH Zurich, Zurich 8092, Switzerland; [eeee]Department of Economics, College of Management Sciences, Michael Okpara University of Agriculture, Umudike 440109, Abia State, Nigeria; [ffff]School of Politics and International Relations, University College Dublin, Dublin 4, Ireland; [gggg]School of Economics and Finance, Victoria University of Wellington, Wellington 6011, New Zealand; [hhhh]Business School, Universidad de los Andes, Bogota 111711, Colombia; [iiii]Universidad de los Andes, Bogota 111711, Colombia; [jjjj]Vancouver School of Economics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [kkkk]Department of Economics, Tilburg University, Tilburg, North Brabant 5037 AB, The Netherlands; [llll]Department of Economics, School of Business and Economics, Freie Universität Berlin, Berlin 14195, Germany; [mmmm]Department of Business Management, Masaryk University, Brno 602 00, Czech Republic; [nnnn]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; [oooo]Department of Ethics, Governance, and Society, School of Business and Economics, Vrije Universiteit Amsterdam, Noord-Holland, Amsterdam 1081HV, The Netherlands; [pppp]Tinbergen Institute, Amsterdam, Noord-Holland 1018WB, The Netherlands; [qqqq]Department of Accountancy, Economics and Finance, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom; [rrrr]Department of Political Science, Université Laval, Québec, QC G1V0A6, Canada; [ssss]laboratoire d'Économie de

Dauphine, Centre Pour la Recherche EconoMique et ses Applications, Paris 75014, France; [tttt]Department of Political and Social Sciences, European University Institute, Fiesole 50014, Italy; [uuuu]Health, Nutrition, and Population Global Practice, World Bank, Washington, DC 20433; [vvvv]Centre for Health Economics, University of York, Heslington YO10 5DD, United Kingdom; [wwww]Business School, Universidad Adolfo Ibañez, Santiago 7910000, Chile; [xxxx]School of Chemical, Materials and Biological Engineering, University of Sheffield, Sheffield S1 3JD, United Kingdom; [yyyy]Institute of Economics, Eötvös Loránd University Centre for Economic and Regional Studies, Budapest 1097, Hungary; [zzzz]Department of Economics, Central European University, Vienna 1100, Austria; [aaaaa]Department of Economics, Deakin University, Burwood, VIC 3125, Australia; [bbbbb]School of Economics, University College Dublin, Dublin D04 F6X4, Ireland; [ccccc]Centre for Business and Economics Research, University of Coimbra, Coimbra 3000-145, Portugal; [ddddd]Behavioral Science, Geary Institute for Public Poliy, Dublin D04 P9C4, Ireland; [eeeee]Department of Law, Economics, and Sociology, "Magna Graecia" University of Catanzaro, Catanzaro 88100, Italy; [fffff]Department of Economics, McMaster University, Hamilton, ON L8S 4M4, Canada; [ggggg]The Canadian Research Data Centre Network, Hamilton, ON L8S4M4, Canada; [hhhhh]Institute for Psychiatry, Psychology & Neuroscience, King's College London, London SE5 9RJ, United Kingdom; [iiiii]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; [jjjjj]Economics, University of California San Diego, La Jolla, California 92092; [kkkkk]Department of Sociology and Human Geography, University of Oslo, Oslo N-0317, Norway; [lllll]School of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom; [mmmmm]University College Dublin Lochlann Quinn School of Business, University College Dublin, Dublin 4, Ireland; [nnnnn]Department of Accounting and Control, University of Lausanne, Lausanne 1015, Switzerland; [ooooo]Mathematics Institute, Leiden University, Leiden, South Holland 2333CA, The Netherlands; [ppppp]Department of Economics and Economic History, Unit of Economic Analysis, Universitat Autònoma de Barcelona, Bellaterra, Cerdanyola de Vallès 08193, Spain; [qqqqq]Department of Economics, Finance, and Legal Studies, University of Alabama, Tuscaloosa, AL 35487; [rrrrr]Institute for Futures Studies, Stockholm 10131, Sweden; [sssss]Department of Economics, University of Toronto, Toronto, ON M5S 3G7, Canada; [ttttt]Computational Social Science, ETH Zürich, Zurich 8057, Switzerland; [uuuuu]Department of Politics and International Relations, University of Oxford, Oxford OX1 3UQ, United Kingdom; [vvvvv]Department of Real Estate Management, Ted Rogers School of Management, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada; [wwwww]School of Psychology, University of Nottingham, Nottingham NG7 2RD, United Kingdom; [xxxxx]Department of Health Sciences and Technology, ETH Zurich, Zurich 8001, Switzerland; [yyyyy]Management School, HEC Liège, University of Liege, Liège 4000, Belgium; [zzzzz]Department of Psychology, University of York, York YO10 5DD, United Kingdom; [aaaaaa]School of Psychology, University of Birmingham, Birmingham B15 2SA, United Kingdom; [bbbbbb]Political Economy Research Institute, University of Massachusetts Amherst, Amherst, MA 01003; [cccccc]Center for Economic and Policy Research, Washington, DC 20009; [dddddd]Research Institute of Industrial Economics, Stockholm 10215, Sweden; [eeeeee]Faculty of Economic Sciences, University of Warsaw, Warsaw 00-927, Poland; [ffffff]Financial Stability Department, Climate Analysis Team, Bank of Canada, Ottawa, ON K1A 0G9, Canada; [gggggg]Bart's Life Sciences, Bart's National Health Services Trust, London E14 5HJ, United Kingdom; [hhhhhh]Queen Mary University of London, London E1 4NS, United Kingdom; [iiiiii]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853; [jjjjjj]Carleton University, Ottawa, ON K1S 5B6; [kkkkkk]University of Oxford, Oxford OX1 2JD, United Kingdom; [llllll]Department of Management and Organization, School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Noord-Holland 1081 HV, Netherlands; [mmmmmm]Department of Economics, University of Freiburg, Freiburg im Breisgau 79085, Germany; [nnnnnn]Department of Clinical Research, University of Basel, Basel 4051, Switzerland; [oooooo]University Hospital Basel, Basel 4031, Switzerland; [pppppp]Department of Sociology, Ludwig Maximilian University of Munich, Munich 80539, Germany; [qqqqqq]Department of Economics, Faculty of Economics and Statistics, University of Innsbruck, Innsbruck 6020, Austria; [rrrrrr]Department of Economics, Turku School of Economics, University of Turku, Turku FI-20014, Finland; [ssssss]Department of Health Economics, Ludwig-Maximilians-Universität Munich, Munich DE-80539, Germany; [tttttt]School of Business and Economics, Freie Universität Berlin, Berlin 14195, Germany; [uuuuuu]Department of Political Science, University of Toronto, Toronto, ON M5S 3G5, Canada; [vvvvvv]Department of Food, Agricultural and Resource Economics, University of Guelph, Guelph, ON N1G 2W1, Canada; [wwwwww]Department of Economics, Yale University, New Haven, CT 06520-8268; [xxxxxx]Research School of Economics, Australian National University, Canberra, ACT 2600, Australia; [yyyyyy]School of Finance, Renmin University of China, Beijing 100872, China; [zzzzzz]Department of Advanced Social and International Studies, University of Tokyo, Meguro City, Tokyo 153-8902, Japan; [aaaaaaa]Dyson School of Applied Economics and Management, Cornell University, Ithaca, NY 14850; [bbbbbbb]Department of Economics, Knauss School of Business, University of San Diego, San Diego, CA 92110; [ccccccc]Department of Economics and Business Administration, University of Cagliari, Cagliari 09124, Italy; [ddddddd]Department of Economics, School of Administrative and Economic Sciences, Pontificia Universidad Javeriana, Bogotá 110231, Colombia; [eeeeeee]Department of Economics, Durham University, Durham DH1 3LB, United Kingdom; [fffffff]School of Economics and Management, Tilburg University, Tilburg 5000 LE, The Netherlands; [ggggggg]Applied Economics, University of Minnesota, Saint Paul, MN 55108; [hhhhhhh]Department of Economics, School of Finance, Economics and Government, Universidad Escuela Finanzas Economía y Gobierno, Antioquia, Medellín 050022, Colombia; [iiiiiii]Samuel Curtis Johnson School of Business, Cornell University, Ithaca, NY 14853; [jjjjjjj]School of Economics, University of Sheffield, Sheffield S10 2TU, United Kingdom; [kkkkkkk]School of Economics and Business, University of Halle, Halle (Saale) 06108, Germany; [lllllll]Department of Marketing, Marshall School of Business, University of Southern California, Los Angeles, CA 90089-1424; [mmmmmmm]Equalis Capital, Paris 75116, France; [nnnnnnn]Department of Economics, Rice University, Houston, TX 77005; [ooooooo]Institute for Financial Management and Research Graduate School of Business, Krea University, Sri City, Andhra Pradesh 517646, India; [ppppppp]Department of Psychology, Dilla University, Dilla, South Ethiopia 419, Ethiopia; [qqqqqqq]Center PhD Students, Research Group: Econometrics, Tilburg School of Economics and Management, Tilburg university, Tilburg 5037 AB, Netherlands; [rrrrrrr]Department of Economics, University of Ottawa, Ottawa, ON K1N 6N5, Canada; [sssssss]Department of Economics, University of Calgary, Calgary, AB T2N 1N4, Canada; [ttttttt]Resources for the Future, Washington, DC 20036; [uuuuuuu]Research Group: Econometrics, Tilburg School of Economics and Management, Tilburg University, Tilburg 5037 AB, The Netherlands; [vvvvvvv]Department of Political Science, University of Montreal, Montreal, QC H3T 1J4, Canada; [wwwwwww]Institut de valorisation des données, Montreal, QC H3N 1V5, Canada; [xxxxxxx]Department of Medicine and Surgery, University of Milano-Bicocca, Milan 20126, Italy; [yyyyyyy]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan 20126, Italy; [zzzzzzz]Department of Health Sciences, University of York, York YO10 5DD, United Kingdom; [aaaaaaaa]Climate and Development Policy Division, RWI - Leibniz Institute for Economic Research, Berlin 10115, Germany; [bbbbbbbb]School of Public Policy and Administration, Carleton University, Ottawa, ON K1S 5B6, Canada; [cccccccc]Institute of Psychology, Cardinal Stefan Wyszynski University, Warsaw 01-938, Poland; [dddddddd]Currency Department, Bank of Canada, Ottawa, ON K1A 0G9, Canada; [eeeeeeee]Department of Agricultural Economics and Rural Development, University of Göttingen, Göttingen, Niedersachsen 37083, Germany; [ffffffff]Instituto de Estudios Superiores de la Empresa Business School, University of Navarra, Barcelona 08034, Spain; [gggggggg]Universidad de Navarra, Pamplona 31009, Spain; [hhhhhhhh]Department of Economics, Dedman College of Humanities and Sciences, Southern Methodist University, Dallas, TX 75275-0496; [iiiiiiii]SKEMA Business School, Groupe de Recherche en Droit, Économie et Gestion, Université Côte d'Azur, Lille 59777, France; [jjjjjjjj]Institute for Public Management and Governance, Vienna University of Economics and Business, Vienna 1020, Austria; [kkkkkkkk]Department of Finance, Rotterdam School of Management, Erasmus University Rotterdam, Rotterdam, Zuid Holland 3012 CC The Netherlands; [llllllll]Institute of Psychology, Universität Osnabrück, Osnabrück 49078, Germany; [mmmmmmmm]Department of Psychology, University of Houston, Houston, TX 77204; [nnnnnnnn]Department of Economics, Business and Finance, Lake Forest College, Lake Forest, IL 60048; [oooooooo]Department of Marketing, Vienna University of Economics and Business, Vienna 1020, Austria; [pppppppp]Department of Economics, Bielefeld University, Bielefeld 33615, Germany; [qqqqqqqq]Department of Economics, University at Albany, State University of New York, Albany, NY 12222; [rrrrrrrr]School of Information, University of Michigan, Ann Arbor, MI 48109; [ssssssss]The Unjournal, Camden, DE 19934; [tttttttt]Finnish Flagship Inequalities, Interventions, and New Welfare State Flagship Research Centre, University of Turku, Turku 20014, Finland; [uuuuuuuu]Applied Economics and Management, Cornell University, Ithaca, NY 14853; [vvvvvvvv]Department of Economics, City St George's, University of London, London EC1V 0HB, United Kingdom; [wwwwwwww]Department of Psychology, University of Milano-Bicocca, Milan 20126, Italy; [xxxxxxxx]School of Economics, University of Nottingham, Nottingham NG7 2RD, United Kingdom; [yyyyyyyy]InsIDE Lab, Dipartimento di Economia, Università degli Studi dell'Insubria, Varese 21100, Italy; [zzzzzzzz]Department of Economics, University of Essex, Colchester CO4 3SQ, United Kingdom; [aaaaaaaaa]Data and Digital Services Department, Bank of Canada, Ottawa, ON K1A 0G9, Canada; [bbbbbbbbb]Systems and Computer Engineering Department, Carleton University, Ottawa, ON K1S 5B6, Canada; [ccccccccc]Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge CB2 1SB, United Kingdom; [ddddddddd]Department of Economics, University of California, Davis, CA 95616; [eeeeeeeee]Institute for Business Administration, University of Regensburg, Regensburg, Bavaria 93053 Germany; [fffffffff]Department of Linguistics, Ghent University, Gent 9000, Belgium; [ggggggggg]Department of Linguistics and English Language, University of Manchester, Manchester M13 9PL, United Kingdom; [hhhhhhhhh]Department of Marketing, Tilburg School of Economics and Management, Tilburg University, Tilburg 5037 AB, The Netherlands; [iiiiiiiii]School of Public Finance and Taxation, Dongbei University of Finance and Economics, Dalian 116025, China; [jjjjjjjjj]Department of Marketing, Clemson University, Clemson, SC 29634; [kkkkkkkkk]Department of Economics, University of Birmingham, Birmingham B15 2TT, United Kingdom; [lllllllll]School of Information, Journalism and Communication, University of Sheffield, Sheffield S10 2AH, United Kingdom; [mmmmmmmmm]Department of Government, Cornell University, Ithaca, NY 14853; [nnnnnnnnn]Agri-Science Queensland, Department of Primary Industries, Brisbane, QLD 4102, Australia; [ooooooooo]Asia and Pacific Department, International Monetary Fund, Washington, DC 20431; [ppppppppp]Department of Psychology, New York University, New York, NY 10003; [qqqqqqqqq]School of Natural and Social Sciences, Scotland's Rural College, Edinburgh EH9 3JG, United Kingdom; [rrrrrrrrr]Chair of Agricultural Production and Resource Economics, Technical University of Munich, Freising 85354, Germany; [sssssssss]The Institute for Food and Resource Economics, University of Bonn, Bonn 53115, Germany; [ttttttttt]Department of Special Needs Education, Centre for Research on Equality in Education, University of Oslo, Oslo 0318, Norway; [uuuuuuuuu]World Bank, Washington, DC 20433; [vvvvvvvvv]Birkbeck Business School, Birkbeck, University of London, London WC1E 7JL, United Kingdom; [wwwwwwwww]Department of Economics, Vanderbilt University, Nashville, TN 37235-1819; [xxxxxxxxx]Division of Psychology & Mental Health, School of Health Sciences, University of Manchester, Manchester M13 9PL, United Kingdom; [yyyyyyyyy]Center for Health and Wellbeing, Princeton University, Princeton, NJ 08544; [zzzzzzzzz]Department of Economics, College of Staten Island, Staten Island, NY 10314; [aaaaaaaaaa]City University of New York, New York, NY 10017; [bbbbbbbbbb]Department Economics, Law, and Society, Ecole upérieure des sciences commerciales School of Management, Angers 49003, France; [cccccccccc]Department of Economics, University of California, Davis, CA 95616; [dddddddddd]Stockholm Business School, Stockholm University, Stockholm 10691, Sweden; [eeeeeeeeee]Leibniz Institute for Financial Research Sustainable Architecture for Finance in Europe, Frankfurt, Hesse 60323, Germany; [ffffffffff]Department of Economics, National University of Singapore, Singapore 117570, Singapore; [gggggggggg]Collegium of World Economy, Szkoła Główna Handlowa Warsaw School of Economics, Warsaw 02-554, Poland; [hhhhhhhhhh]Department of Political Science, Stanford University, Stanford, CA 94305; [iiiiiiiiii]Department of Engineering Sciences & Applied Mathematics, Northwestern University, Evanston, IL 60201; [jjjjjjjjjj]Real Estate Economics, National Chengchi University, Taipei 11605, Taiwan; [kkkkkkkkkk]Aalto University, Espoo 02150, Finland; [llllllllll]Department of Accounting, Tilburg School of Economics and Management, Tilburg University, Tilburg 5037 AB, The Netherlands; [mmmmmmmmmm]Department of Economics, Faculty of Arts and Social Sciences, National University of Singapore, Singapore 117570, Singapore; [nnnnnnnnnn]School of Economics, Faculty of Buisness Economics and Law, University of Queensland, Brisbane, St Lucia, QLD 4072, Australia; [oooooooooo]Department of Psychology, Faculty of Science and Technology, Fylde College, Lancaster University, Lancaster LA1 4YF, United Kingdom; [pppppppppp]School of Biosciences, University of Sheffield, Sheffield S10 2TN, United Kingdom; [qqqqqqqqqq]Christ Church College, University of Oxford, Oxfordshire OX1 1DP, United Kingdom; [rrrrrrrrrr]Economics Experimental Lab, Nanjing Audit University, Nanjing 210017, China; [ssssssssss]Department of Finance, HEC, University of Lausanne, Lausanne 1015, Switzerland; [tttttttttt]Swiss Finance Institute, Lausanne 1015, Switzerland; [uuuuuuuuuu]Department of Marketing, Rotman School of Management, University of Toronto, Toronto, ON M5S 1A1, Canada; [vvvvvvvvvv]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027; and [wwwwwwwwww]Department of Economics, Penn State University, University Park, PA 16802

1. A. Brodeur *et al*., Promoting reproducibility and replicability in political science. *Res. Polit.* **11**, 20531680241233439 (2024).
2. D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, V. Stodden, Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* **11**, 8–18 (2008).
3. M. Fišar *et al*., Reproducibility in management science. *Manag. Sci.* **70**, 1343–1356 (2024).
4. P. Gertler, S. Galiani, M. Romero, How to make replication the norm. *Nature* **554**, 417–419 (2018).
5. S. N. Goodman, D. Fanelli, J. P. Ioannidis, What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12-341ps12 (2016).
6. M. Milkowski, W. M. Hensel, M. Hohol, Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *J. Comput. Neurosci.* **45**, 163–172 (2018).
7. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (National Academies Press, 2019).
8. C. Pérignon, K. Gadouche, C. Hurlin, R. Silberman, E. Debonnel, Certify reproducibility with confidential data. *Science* **365**, 127–128 (2019).
9. L. Vilhuber, Report by the AEA data editor. *AEA Pap. Proc.* **112**, 813–823 (2022).
10. A. Brodeur *et al*., Reproducibility and robustness of economics and political science research. *Nature* **652**, 151–156 (2026).
11. A. C. Chang, P. Li, Is economics research replicable? Sixty published papers from thirteen journals say "often not". *Crit. Finance Rev.* **11**, 185–206 (2022).
12. 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.. *Psychol. Sci.* **34**, 512–522 (2023).
13. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
14. P. Obels, D. Lakens, N. A. Coles, J. Gottfried, S. A. Green, Analysis of open data and computational reproducibility in registered reports in psychology. *Adv. Methods Pract. Psychol. Sci.* **3**, 229–237 (2020).
15. C. Pérignon *et al*., Computational reproducibility in finance: Evidence from 1,000 tests. *Rev. Financial Stud.* **37**, 3558–3593 (2024).
16. V. Stodden, J. Seiler, Z. Ma, An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2584–2589 (2018).
17. B. Wood, R. Müller, A. Brown, Push button replication: Is impact evaluation evidence for international development verifiable? *PloS one* **13**, e0209416 (2018).
18. J. E. Colliard, C. Hurlin, C. Perignon, The Economics of Computational Reproducibility (2022). https://ssrn.com/abstract=3418896.
19. A. Hrycyshyn, H. Eassom, "ExplanAItions: An AI study" (Tech. Rep. 1, John Wiley & Sons, Hoboken, NJ, 2025).
20. M. Vaccaro, A. Almaatouq, T. Malone, When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nat. Hum. Behav.* **8**, 1–11 (2024).
21. 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.
22. 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* (PKP Publishing Services Network, 2021), vol. 35, pp. 11405–11414.
23. E. Bondi *et al*., "Role of human-AI interaction in selective prediction" in *Proceedings of the AAAI Conference on Artificial Intelligence* (PKP Publishing Services Network, 2022), vol. 36, pp. 5286–5294.
24. Á. A. Cabrera, A. Perer, J. I. Hong, Improving human-AI collaboration with descriptions of AI behavior. *Proc. ACM Hum. Comput. Interact.* **7**, 1–21 (2023).
25. E. Goh *et al*., Influence of a large language model on diagnostic reasoning: A randomized clinical vignette study. medRxiv [Preprint] (2024). https://doi.org/10.1101/2024.03.12.24303785 (Accessed 20 February 2026).
26. B. Koepnick *et al*., De novo protein design by citizen scientists. *Nature* **570**, 390–394 (2019).
27. H. Liu, V. Lai, C. Tan, Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proc. ACM Hum. Comput. Interact.* **5**, 1–45 (2021).
28. H. Mozannar *et al*., Effective human-AI teams via learned natural language rules and onboarding. *Adv. Neural Inf. Process. Syst.* **36**, e01007 (2024).
29. S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
30. C. Reverberi *et al*., Experimental evidence of effective human-AI collaboration in medical decision-making. *Sci. Rep.* **12**, 14952 (2022).
31. 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.
32. M. H. Tessler *et al*., AI can help humans find common ground in democratic deliberation. *Science* **386**, eadq2852 (2024).
33. M. Vaccaro, J. Waldo, The effects of mixing machine learning and human judgment. *Commun. ACM* **62**, 104–110 (2019).
34. B. Wilder, E. Horvitz, E. Kamar, "Learning to complement humans" in *Proceedings of the 29th International Joint Conference on Artificial Intelligence* (ACM Digital Library, 2020), pp. 1526–1533.
35. Cherry Bekaert LLP, Report of independent auditor. *Am. Econ. Rev*. **112**, 2083–2098 (2022).
36. 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. *Proc. ACM Hum. Comput. Interact.* **5**, 1–21 (2021).
37. L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making? *Int. J. Hum. Comput. Stud.* **51**, 991–1006 (1999).
38. E. Brynjolfsson, D. Li, L. Raymond, Generative AI at work. *Q. J. Econ.* **140**, 889–942 (2025).
39. M. Del Giudice, S. W. Gangestad, A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Adv. Methods Pract. Psychol. Sci.* **4**, 2515245920954925 (2021).
40. X. Lu, H. White, Robustness checks and robustness tests in applied economics. *J. Econom.* **178**, 194–206 (2014).
41. M. B. Nuijten, "Assessing and improving robustness of psychological research findings in four steps" in *Avoiding Questionable Research Practices in Applied Psychology* (Springer, 2022), pp. 379–400.
42. V. Lai, H. Liu, C. Tan,"'why is' chicago'deceptive?' Towards building model-driven tutorials for human" in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (ACM Digital Library, 2020), pp. 1–13.
43. 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* (ACM Digital Library, 2020), pp. 295–305.
44. H. Vasconcelos *et al*., Explanations can reduce overreliance on AI systems during decision-making. *Proc. ACM Hum. Comput. Interact.* **7**, 1–38 (2023).
45. 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. *Int. J. Inf. Manag.* **71**, 102642 (2023).
46. 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. *J. Assoc. Inf. Sci. Technol.* **74**, 570–581 (2023).
47. N. Wadhwa *et al*., Core: Resolving code quality issues using LLMs. *Proc. ACM Softw. Eng.* **1**, 789–811 (2024).
48. 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* (ACM Digital Library, 2024), pp. 683–685.
49. 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* (ACM Digital Library, 2024), pp. 1–13.
50. A. Brodeur *et al.*, AI Replication Games. https://github.com/I4Replication/AI-Games. GitHub. Accessed 2 May 2026.