Preregister Experiments with AI Agents

Michelle Vaccaro

MIT Institute for Data, Systems, and Society

People are treating AI agents as participants in experiments

Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies

Gati Aher [1] Rosa I. Arriaga [2] Adam Tauman Kalai [3]

Evaluating the Moral Beliefs Encoded in LLMs

Warning: This paper contains moral scenarios which are controversial and offensive in nature.

Nino Scherrer *[1], Claudia Shi *[1,2], Amir Feder [2], and David M. Blei [2]
[1] FAR AI, [2] Columbia University

EVALUATING LANGUAGE MODEL AGENCY THROUGH NEGOTIATIONS

Tim R. Davidson[†*] Veniamin Veselovsky[†*] Martin Josifoski[†] Maxime Peyrard[‡]
Antoine Bosselut[†] Michal Kosinski[§] Robert West[†]
[†]EPFL, [‡]UGA, CNRS, LIG, [§]Stanford University

Evaluating and Inducing Personality in Pre-trained Language Models

Guangyuan Jiang[1,2,*] jgy@stu.pku.edu.cn
Manjie Xu[1,*] manjietsu@gmail.com
Song-Chun Zhu[1,3] s.c.zhu@pku.edu.cn
Wenjuan Han[4,✉] wjhan@bjtu.edu.cn
Chi Zhang[3,✉] zhangchi@bigai.ai
Yixin Zhu[1,✉] yixin.zhu@pku.edu.cn

MACHINE PSYCHOLOGY

Thilo Hagendorff* University of Stuttgart
Ishita Dasgupta* Google DeepMind
Marcel Binz† Helmholtz Institute for Human-Centered AI
Stephanie C.Y. Chan† Google DeepMind
Andrew Lampinen† Google DeepMind
Jane X. Wang† Google DeepMind
Zeynep Akata TU Munich
Eric Schulz Helmholtz Institute for Human-Centered AI

Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias

Itay Itzhak[1], Gabriel Stanovsky[2], Nir Rosenfeld[1], Yonatan Belinkov[1]
[1]Technion – Israel Institute of Technology, Israel
[2]School of Computer Science and Engineering, The Hebrew University of Jerusalem
itaylitzhak@gmail.com,
{nirr, belinkov}@technion.ac.il, gabriel.stanovsky@mail.huji.ac.il

Experiments with AI agents exacerbate vulnerabilities that exist in experiments with human subjects

› Psychol Sci. 2011 Nov;22(11):1359-66. doi: 10.1177/0956797611417632. Epub 2011 Oct 17.

False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant

Joseph P Simmons [1], Leif D Nelson, Uri Simonsohn

Affiliations + expand

PMID: 22006061 DOI: 10.1177/0956797611417632

Abstract

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

› Psychol Sci. 2012 May 1;23(5):524-32. doi: 10.1177/0956797611430953. Epub 2012 Apr 16.

Measuring the prevalence of questionable research practices with incentives for truth telling

Leslie K John [1], George Loewenstein, Drazen Prelec

Affiliations + expand

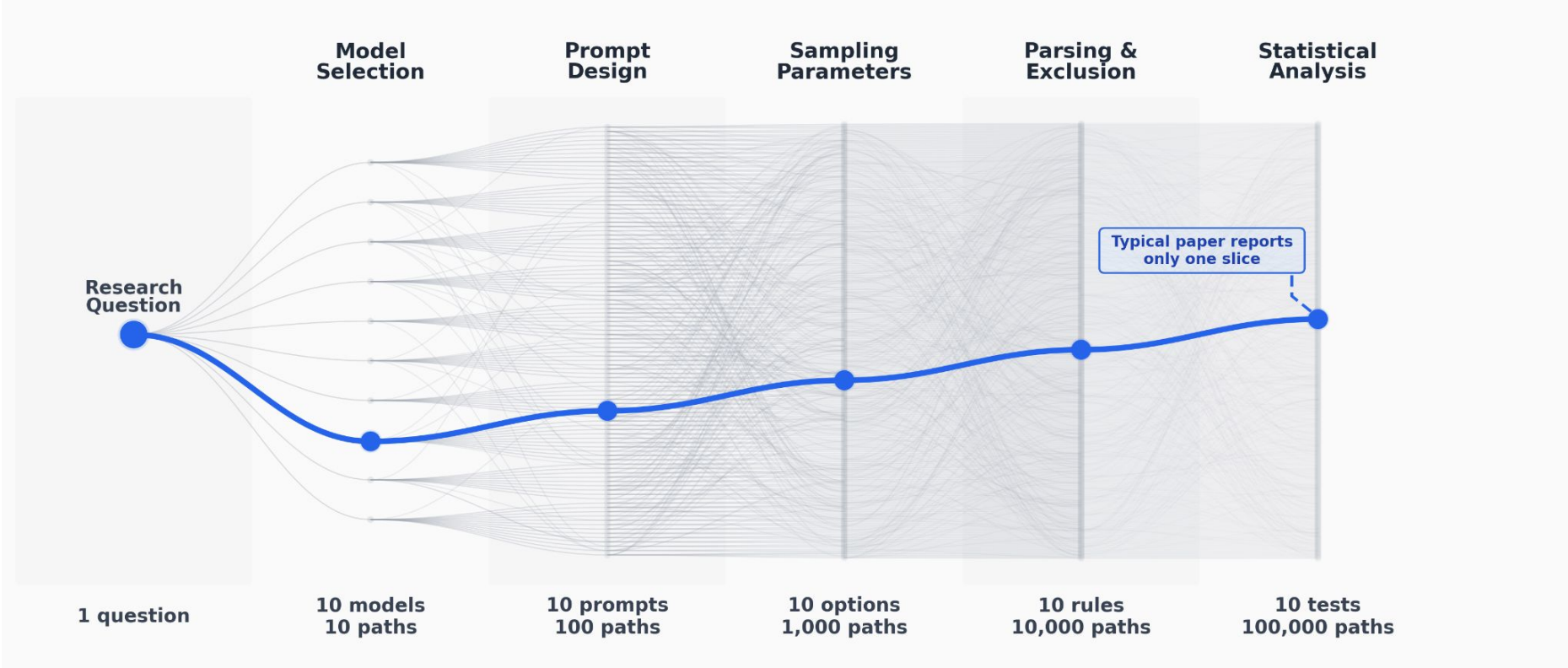
PMID: 22508865 DOI: 10.1177/0956797611430953

Abstract

Cases of clear scientific misconduct have received significant media attention recently, but less flagrantly questionable research practices may be more prevalent and, ultimately, more damaging to the academic enterprise. Using an anonymous elicitation format supplemented by incentives for honest reporting, we surveyed over 2,000 psychologists about their involvement in questionable research practices. The impact of truth-telling incentives on self-admissions of questionable research practices was positive, and this impact was greater for practices that respondents judged to be less defensible. Combining three different estimation methods, we found that the percentage of respondents who have engaged in questionable practices was surprisingly high. This finding suggests that some questionable practices may constitute the prevailing research norm.

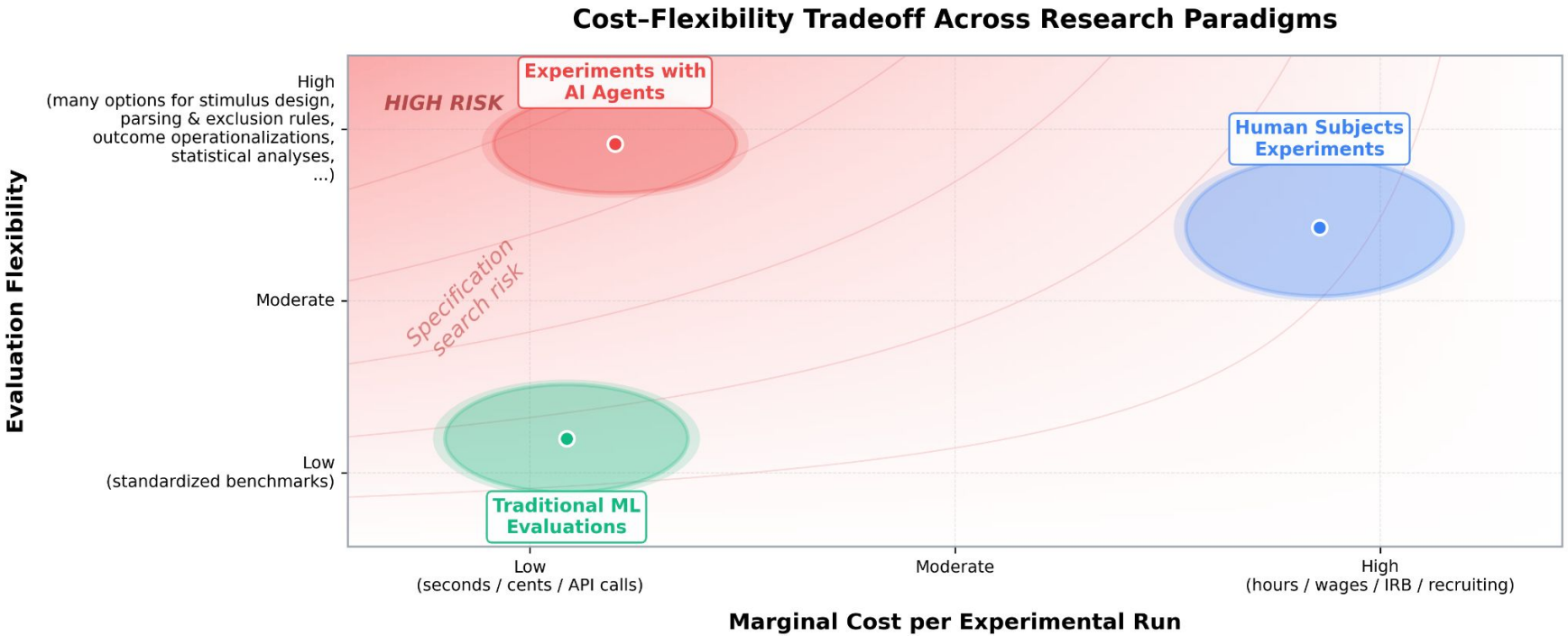
HARKing, *p*-hacking, design-hacking, selective reporting…

Experiments with AI agents exacerbate vulnerabilities that exist in experiments with human subjects



Combinatorial explosion of researcher degrees of freedom

Experiments with AI agents exacerbate vulnerabilities that exist in experiments with human subjects



Low experimental cost, high evaluation flexibility = high specification search risk

We should preregister experiments with AI agents to address these issues

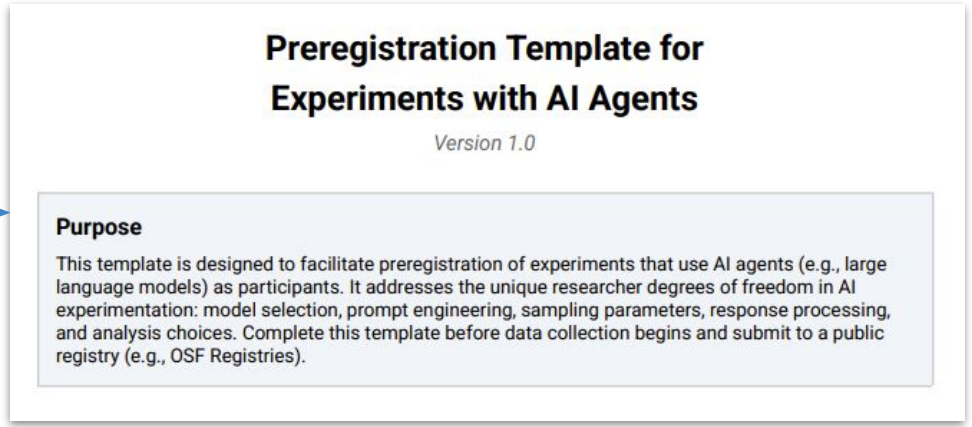
RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

The preregistration revolution

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor Authors Info & Affiliations

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved August 28, 2017 (received for review June 15, 2017)

March 12, 2018 | 115 (11) 2600-2606 | https://doi.org/10.1073/pnas.1708274114



Preregistration Template for Experiments with AI Agents

Version 1.0

Purpose

This template is designed to facilitate preregistration of experiments that use AI agents (e.g., large language models) as participants. It addresses the unique researcher degrees of freedom in AI experimentation: model selection, prompt engineering, sampling parameters, response processing, and analysis choices. Complete this template before data collection begins and submit to a public registry (e.g., OSF Registries).

Extending the logic of traditional preregistration for the new degrees of freedom that characterize AI agent research

Thank you!

Questions? Comments? Suggestions?

Please leave a comment or reach out to me!