SILICON REVIEWERS: SIMULATING EXPERT PANELS TO TEST FUNDING MECHANISMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inter-reviewer agreement in peer review hovers around $\kappa = 0.17$, barely above chance; a quarter of funding decisions would reverse with different reviewers. Yet we lack methods for testing whether proposed reforms improve outcomes, because experiments on live funding are expensive and ethically fraught. We propose the *grant simulation machine*: AI agents trained on historical evaluation data to replicate expert panel behavior, validated against real decisions, then used to test counterfactual funding mechanisms in silico. Drawing on AI-based human simulation (85% accuracy modeling individual behavior) and a partnership providing complete evaluation-to-outcome data, we outline a framework for comparing golden tickets, partial lotteries, and alternative panel compositions. The approach faces real limitations: AI may not capture emergent group dynamics, and simulated outcomes cannot substitute for actual scientific productivity. But it offers something currently lacking: a low-stakes laboratory for institutional innovation.

1 INTRODUCTION

Twenty-five percent of NSF funding decisions would reverse with different reviewers (Cole et al., 1981). Meta-analyses find mean inter-rater reliability of $\kappa = 0.17$ across peer review contexts (Bornmann et al., 2010), barely better than chance. Yet despite decades of documented problems, we run essentially the same evaluation machinery our grandparents used.

The obstacle isn't ignorance about alternatives. Golden ticket mechanisms (letting each panelist champion one high-risk proposal), partial lotteries (randomizing among proposals above a quality threshold), and diverse panel compositions have all been proposed (Chawla, 2023; Liu et al., 2020). The obstacle is evidence. Experiments on live funding raise ethical concerns (some applicants get disadvantaged), practical barriers (funders cannot easily randomize core processes), and statistical challenges (small samples, years to outcome).

As AI capabilities approach AGI, this problem intensifies: systems that can generate unlimited research proposals strain evaluation infrastructure designed for human-paced science. The fundamental challenge shifts from allocating funding among proposals to allocating *verification effort* among AI-generated discoveries. We propose testing these mechanisms in simulation first. Park et al. (2024) achieved 85% accuracy simulating real individuals' responses across diverse contexts; Argyle et al. (2023) showed demographically conditioned AI matches actual survey distributions. If AI can simulate voters and experimental subjects, can it simulate grant reviewers?

Through a partnership with Novo Nordisk, we have access to complete evaluation records: proposals, reviewer backgrounds, scores, rationales, decisions, and eventual research outcomes. This ground truth enables both calibration (training AI agents to match reviewer behavior) and validation (testing whether simulated mechanisms predict improved outcomes). The question isn't whether peer review needs reform; the evidence is overwhelming. It's whether we can build infrastructure to know which reforms work.

2 THE PROBLEM: PEER REVIEW'S KNOWN FAILURES

2.1 LOW AGREEMENT, WEAK PREDICTION

The evidence is consistent. Bornmann et al. (2010) meta-analyzed inter-reviewer agreement and found mean $\kappa = 0.17$, classified as "slight" agreement. Pier et al. (2018) had 43 reviewers evaluate 25 NIH proposals; proposals ranked top-quartile by some reviewers landed bottom-quartile for others. Cole et al. (1981) estimated 25% of funded proposals would be rejected, and 25% of rejected proposals funded, with different equally-qualified reviewers.

Do noisy scores still predict success? Fang et al. (2016) found NIH peer review scores predict productivity with AUC = 0.54, barely above random (0.50).

2.2 SYSTEMATIC BIAS

Beyond noise, peer review exhibits systematic biases. Tomkins et al. (2017) found single-blind reviewers favored famous authors (63% higher acceptance odds) and prestigious institutions (58% higher odds), independent of quality. Bol et al. (2018) documented the Matthew effect: researchers who narrowly win early grants accumulate twice as much funding over eight years as equally-qualified narrow losers. The system amplifies initial noise into lasting inequality.

2.3 THE REFORM BOTTLENECK

Running experiments on live funding is hard. Consider testing a partial lottery where proposals above a quality threshold get randomly selected rather than rank-ordered. Such a system might reduce noise and bias while maintaining minimum quality. But implementing it means some researchers who would have been funded under the old system lose out. Ethics committees reasonably question whether "we're running an experiment" justifies individual harm.

Sample sizes compound the problem. Detecting meaningful differences requires many funding cycles spanning years; by the time results emerge, institutional memory has faded.

3 AI-SIMULATED EXPERT PANELS

3.1 THE CASE FOR SIMULATION

What if we could test funding mechanisms without touching real applicants? Train AI agents to behave like real reviewers, validate that they reproduce actual panel decisions, then explore counterfactual designs. Wrong mechanisms fail cheaply in simulation rather than expensively in practice.

Horton (2023) showed AI reproduces behavioral economics findings: anchoring, framing effects, social preferences. Argyle et al. (2023) demonstrated AI produces "silicon samples" matching demographically-defined subgroups. Park et al. (2023) showed emergent social dynamics in simulated agents. Most impressively, Park et al. (2024) achieved 85% accuracy predicting specific individuals' responses across diverse tasks.

3.2 FROM INDIVIDUALS TO PANELS

Grant review involves individual judgment and group dynamics. Can AI capture deliberation: status hierarchies, persuasion, coalition formation? The honest answer: partially. We focus first on what's tractable: validating agents against individual scores, testing aggregation mechanisms that don't depend on deliberation. Group dynamics remain an open challenge we discuss explicitly.

3.3 VALIDATION STRATEGY

Simulation is only useful if it's accurate. Our validation strategy has three components:

Individual calibration. For each reviewer in our historical data, we train an AI agent on their past evaluations: proposals reviewed, scores assigned, written comments. We then test whether the

agent predicts held-out evaluations. The target isn't perfection; it's correlation strong enough that mechanism comparisons remain meaningful.

Panel-level validation. Even if individual predictions have noise, panel-level outcomes might be more stable. We simulate entire funding rounds and compare predicted funding decisions to actual decisions. What fraction of grants are correctly classified?

Outcome linkage. The deepest validation connects simulated decisions to real outcomes. If a simulated mechanism would have funded different proposals, would those proposals have produced better research? This requires linking evaluation data to productivity measures: publications, citations, patents. Our partnership provides exactly this linkage.

4 The Simulation Framework

4.1 Data and Architecture

Through our partnership with Novo Nordisk, we access: full proposal texts, reviewer profiles and histories, numerical scores with written critiques, deliberation summaries, funding decisions, and subsequent outcomes (publications, citations, patents, career trajectories). This complete pipeline, rare because most funders track decisions but not outcomes, enables our validation strategy.

Each simulated reviewer is an AI agent conditioned on static features (expertise, career stage, institution), dynamic features (specific proposals, panel composition), and behavioral patterns learned from historical evaluations (scoring tendencies, innovation preferences). Agents produce both scores and rationales.

4.2 Mechanisms to Test

Golden tickets. Each panelist designates one proposal for automatic advancement regardless of aggregate score. NSF has experimented with versions (Chawla, 2023). Does it surface high-variance proposals? Introduce new biases?

Partial lotteries. Proposals above a threshold enter a lottery rather than being rank-ordered. Liu et al. (2020) found 63% researcher support in New Zealand. Does randomization reduce bias while maintaining quality?

Panel composition. Compare expert-heavy panels against broader stakeholder panels. Which better predicts eventual success?

Deliberation structure. Compare independent scoring with aggregation versus discussion-based formats. When does social influence help versus hurt?

For each mechanism, we simulate many funding rounds and link to outcome data. The mechanism that best predicts success wins.

5 Related Work

Peer Review Reliability and Bias. Decades of research document peer review's limitations. Bornmann et al. (2010) meta-analyzed inter-reviewer agreement at $\kappa = 0.17$ ("slight"). Cole et al. (1981) showed 25% of NSF decisions would reverse with different reviewers. Fang et al. (2016) found NIH scores predict productivity at AUC = 0.54, barely above chance. Systematic biases favor prestigious institutions and famous authors (Tomkins et al., 2017), with the Matthew effect amplifying early noise into lasting inequality (Bol et al., 2018).

Funding Mechanism Alternatives. Proposed reforms include golden tickets for high-risk proposals (Chawla, 2023), partial lotteries randomizing among quality-threshold proposals (Liu et al., 2020), and distributed peer review. Recent work provides extended taxonomies of lottery mechanisms and evaluates their fairness properties. What's missing is empirical evidence: testing mechanisms on live funding is ethically fraught and statistically underpowered.

AI Simulation of Human Behavior. Horton (2023) showed AI reproduces behavioral economics findings; Argyle et al. (2023) demonstrated "silicon samples" matching demographic subgroups.

Park et al. (2024) achieved 85% accuracy simulating specific individuals. Tranchero et al. (2024) proposed Generative AI-Based Experimentation (GABE) for in silico organizational research, showing LLMs can replicate human experiments at lower cost while enabling theory extension. However, Gao et al. (2025) found AI fails in strategic games, and recent work warns that different defensible choices in AI simulation can yield contradictory results.

AI in Peer Review. Current AI applications focus on reviewer matching and plagiarism detection. NeurIPS, ICML, and ICLR have implemented reproducibility initiatives. Our contribution differs: rather than automating peer review, we propose simulating it to test counterfactual mechanisms before deployment.

6 LIMITATIONS

We take seriously what could go wrong:

Emergent dynamics. Panel deliberation involves status hierarchies, persuasion cascades, and groupthink that may not emerge from aggregated individual behavior.

Distribution shift. Agents trained on historical data may not generalize to novel proposals or emerging fields.

Outcome measurement. "Research success" is contested; publications and citations miss societal impact and valuable negative results.

Goodhart's Law. If funders adopt simulation-optimized mechanisms, researchers may game them.

AI limitations. Gao et al. (2025) found AI fails in strategic games; grant review involves strategic elements that may exceed current AI capabilities.

We position simulation as complement, not replacement, for real-world experimentation. It can eliminate bad mechanisms, prioritize promising ones for field trials, and generate testable hypotheses, but it cannot definitively prove a mechanism will work in deployment.

7 CONCLUSION

Science funding allocates billions annually through processes we know are noisy ($\kappa = 0.17$), biased, and weakly predictive (AUC = 0.54). Proposed reforms languish because testing them on live funding disrupts real careers. The simulation approach offers an alternative: build AI agents that replicate reviewer behavior, validate against real decisions and outcomes, then compare mechanisms before deployment. The technical foundations exist; AI achieves 85% accuracy simulating individual judgment. What has been missing is data infrastructure linking decisions to ground-truth outcomes, which our partnership provides.

For the P-AGI research community, this work addresses a structural problem in how we allocate resources to AI research itself. As AI capabilities accelerate, the mechanisms that decide which research gets funded become increasingly consequential. If peer review remains noisy and biased, then the trajectory of AI development depends partly on reviewer lottery. Simulation infrastructure could help: testing whether alternative mechanisms better identify transformative research, reduce bias toward incremental work, or surface high-variance ideas that reviewers currently filter out. The same verification bottleneck that plagues AI-generated scientific claims applies here: we generate more funding mechanism proposals than we can empirically test. Simulation offers a partial escape from this bottleneck.

Three questions now require attention. Technically: can we simulate panel dynamics, not just individual scores, and how robust are mechanism comparisons to simulation error? Institutionally: will funders trust simulation-informed recommendations, and will researchers accept AI-designed mechanisms? Philosophically: if AI systems become accurate enough to simulate expert judgment, what remains of the case for human panels at all?

We are not arguing that simulation should replace human judgment. We are arguing that institutional innovation deserves the same empirical foundation we demand of scientific claims. Peer review is an intervention; we should know whether it works.

REFERENCES

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. The matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19):4887–4890, 2018. doi: 10.1073/pnas.1719557115.

Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS ONE*, 5(12):e14331, 2010. doi: 10.1371/journal.pone.0014331.

Dalmeet Singh Chawla. 'golden tickets' on the cards for NSF grant reviewers. *Nature*, 615:383, 2023. doi: 10.1038/d41586-023-00579-z. News article on NSF experiments with golden ticket mechanisms.

Stephen Cole, Jonathan R. Cole, and Gary A. Simon. Chance and consensus in peer review. *Science*, 214(4523):881–886, 1981. doi: 10.1126/science.7302566.

Ferric C. Fang, Anthony Bowen, and Arturo Casadevall. NIH peer review percentile scores are poorly predictive of grant productivity. *eLife*, 5:e13323, 2016. doi: 10.7554/eLife.13323.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24), 2025. doi: 10.1073/pnas.2501660122. Demonstrates limitations of LLM simulation in the 11-20 money request game.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

Mengyao Liu, Vernon Choy, Philip Clarke, Adrian Barnett, Tony Blakely, and Lucy Pomeroy. The acceptability of using a lottery to allocate research funding: A survey of applicants. *Research Integrity and Peer Review*, 5:3, 2020. doi: 10.1186/s41073-019-0089-z.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 2023. doi: 10.1145/3586183.3606763.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint*, 2024. Achieved 85% accuracy simulating 1,052 real individuals.

Elizabeth L. Pier, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J. Nathan, Cecilia E. Ford, and Molly Carnes. Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, 115(12):2952–2957, 2018. doi: 10.1073/pnas.1714379115.

Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017. doi: 10.1073/pnas.1707323114.

Matteo Tranchero, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj. Theorizing with large language models. Working Paper 33033, National Bureau of Economic Research, 2024. URL `http://www.nber.org/papers/w33033`.

A APPENDIX

LLMs were used for drafting assistance during the preparation of this paper. The authors take full responsibility for all content.