Optimal Science Funding

Aydin Mohseni

DEPARTMENT OF PHILOSOPHY
CARNEGIE MELLON UNIVERSITY
INSTITUTE FOR COMPLEX SOCIAL DYNAMICS

The puzzle: Who should grantmakers fund?

a) the people with the strongest track records

b) those with the least current funding

c) something else...

My Research



Metascience

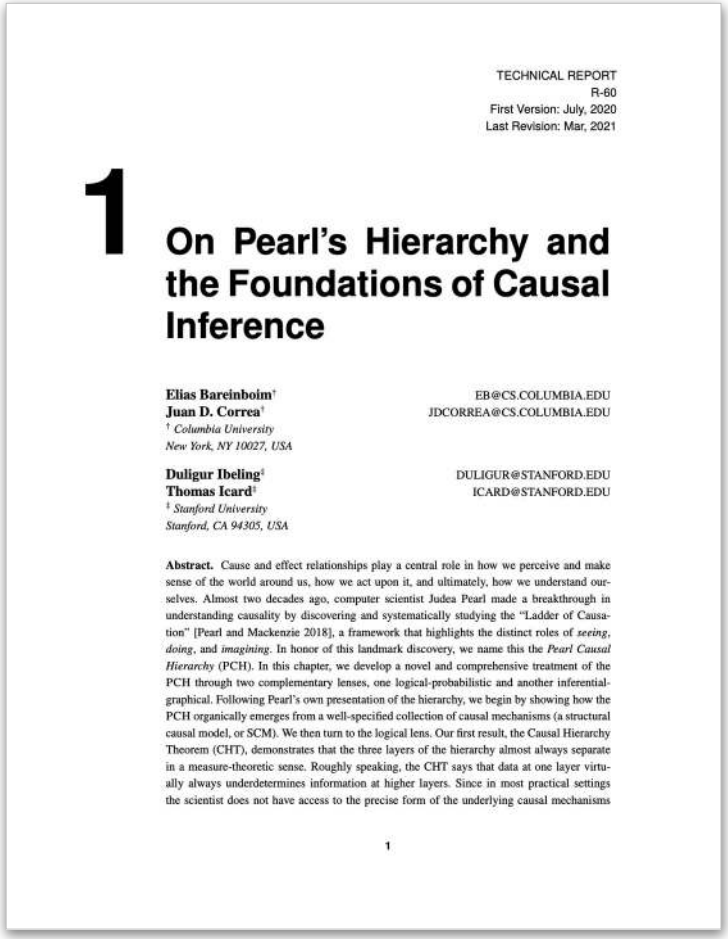


AI Agent Foundations



Agent Foundations

The aim: *produce safe AI using mathematical lenses of Bayesian inference. decision theory, game theory*



TECHNICAL REPORT
R-60
First Version: July, 2020
Last Revision: Mar, 2021

1 On Pearl's Hierarchy and the Foundations of Causal Inference

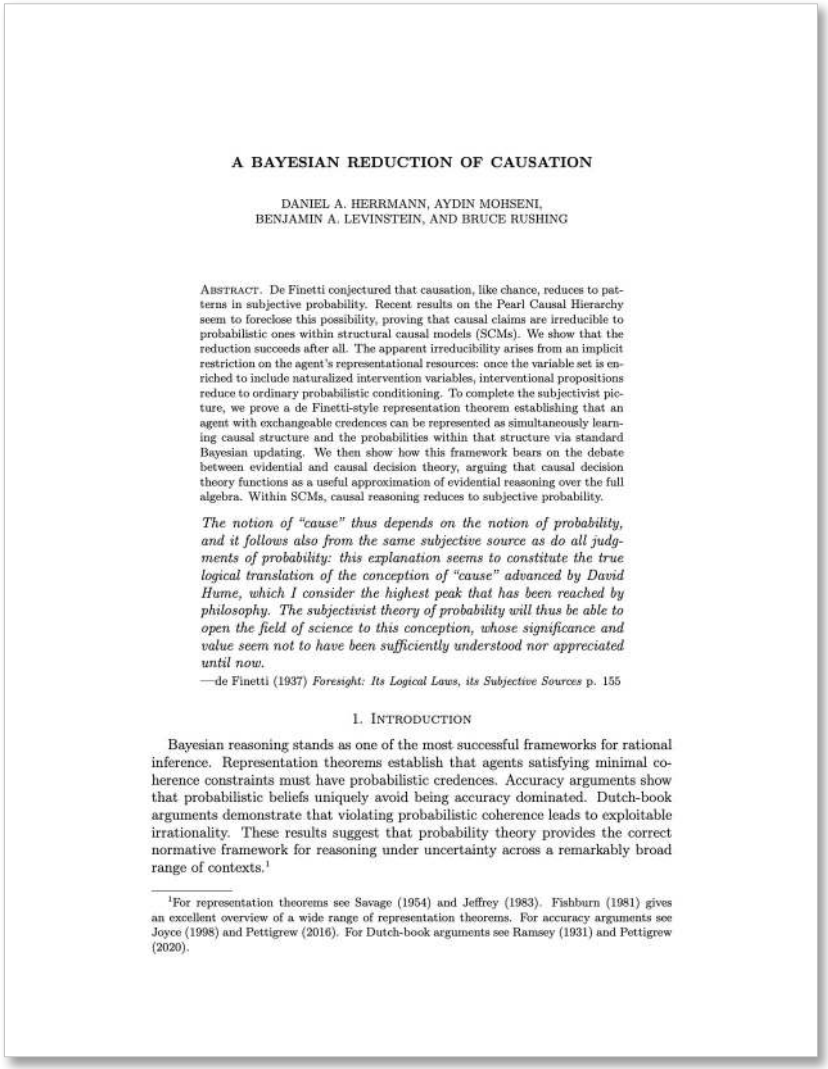
Elias Bareinboim[†] EB@CS.COLUMBIA.EDU
Juan D. Correa[†] JDCORREA@CS.COLUMBIA.EDU
[†] *Columbia University*
New York, NY 10027, USA

Duligur Ibeling[‡] DULIGUR@STANFORD.EDU
Thomas Icard[‡] ICARD@STANFORD.EDU
[‡] *Stanford University*
Stanford, CA 94305, USA

Abstract. Cause and effect relationships play a central role in how we perceive and make sense of the world around us, how we act upon it, and ultimately, how we understand ourselves. Almost two decades ago, computer scientist Judea Pearl made a breakthrough in understanding causality by discovering and systematically studying the "Ladder of Causation" [Pearl and Mackenzie 2018], a framework that highlights the distinct roles of *seeing*, *doing*, and *imagining*. In honor of this landmark discovery, we name this the *Pearl Causal Hierarchy* (PCH). In this chapter, we develop a novel and comprehensive treatment of the PCH through two complementary lenses, one logical-probabilistic and another inferential-graphical. Following Pearl's own presentation of the hierarchy, we begin by showing how the PCH organically emerges from a well-specified collection of causal mechanisms (a structural causal model, or SCM). We then turn to the logical lens. Our first result, the Causal Hierarchy Theorem (CHT), demonstrates that the three layers of the hierarchy almost always separate in a measure-theoretic sense. Roughly speaking, the CHT says that data at one layer virtually always underdetermines information at higher layers. Since in most practical settings the scientist does not have access to the precise form of the underlying causal mechanisms

- "The bulk of human knowledge is organized around causal, not probabilistic relationships, and **the grammar of probability calculus is insufficient for capturing those relationships**..."

—Pearl (2001, p. 31) *Bayesianism and Causality or Why I am Only a Half-Bayesian*



A BAYESIAN REDUCTION OF CAUSATION

DANIEL A. HERRMANN, AYDIN MOHSENI,
BENJAMIN A. LEVINSTEIN, AND BRUCE RUSHING

ABSTRACT. De Finetti conjectured that causation, like chance, reduces to patterns in subjective probability. Recent results on the Pearl Causal Hierarchy seem to foreclose this possibility, proving that causal claims are irreducible to probabilistic ones within structural causal models (SCMs). We show that the reduction succeeds after all. The apparent irreducibility arises from an implicit restriction on the agent's representational resources: once the variable set is enriched to include naturalized intervention variables, interventional propositions reduce to ordinary probabilistic conditioning. To complete the subjectivist picture, we prove a de Finetti-style representation theorem establishing that an agent with exchangeable credences can be represented as simultaneously learning causal structure and the probabilities within that structure via standard Bayesian updating. We then show how this framework bears on the debate between evidential and causal decision theory, arguing that causal decision theory functions as a useful approximation of evidential reasoning over the full algebra. Within SCMs, causal reasoning reduces to subjective probability.

The notion of "cause" thus depends on the notion of probability, and it follows also from the same subjective source as do all judgments of probability: this explanation seems to constitute the true logical translation of the conception of "cause" advanced by David Hume, which I consider the highest peak that has been reached by philosophy. The subjectivist theory of probability will thus be able to open the field of science to this conception, whose significance and value seem not to have been sufficiently understood nor appreciated until now.

—de Finetti (1937) *Foresight: Its Logical Laws, its Subjective Sources* p. 155

1. INTRODUCTION

Bayesian reasoning stands as one of the most successful frameworks for rational inference. Representation theorems establish that agents satisfying minimal coherence constraints must have probabilistic credences. Accuracy arguments show that probabilistic beliefs uniquely avoid being accuracy dominated. Dutch-book arguments demonstrate that violating probabilistic coherence leads to exploitable irrationality. These results suggest that probability theory provides the correct normative framework for reasoning under uncertainty across a remarkably broad range of contexts.[1]

[1]For representation theorems see Savage (1954) and Jeffrey (1983). Fishburn (1981) gives an excellent overview of a wide range of representation theorems. For accuracy arguments see Joyce (1998) and Pettigrew (2016). For Dutch-book arguments see Ramsey (1931) and Pettigrew (2020).

In collaboration with...



Daniel Herrmann

UNC Chapel Hill



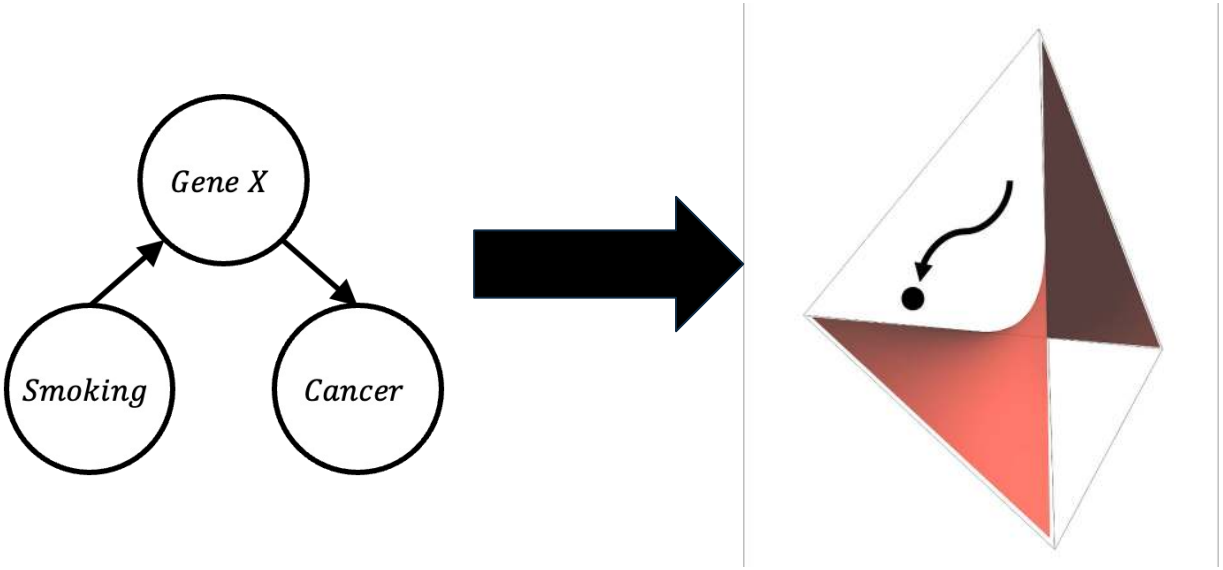
Ben Levinstein

Anthropic



Bruce Rushing

U of Virginia


Gene X
Smoking
Cancer

LEARNING THE VALUE OF VALUE LEARNING

ALEX JOHN LONDON & AYDIN MOHSENI

ABSTRACT. Standard decision frameworks address uncertainty about facts but assume fixed options and values. We extend the Jeffrey–Bolker framework to model refinements in values and prove a value-of-information theorem for axiological refinement. In multi-agent settings, we establish that mutual refinement will characteristically transform zero-sum games into positive-sum interactions and yield Pareto-improvements in Nash bargaining. These results show that a framework of rational choice can be extended to model value refinement. By unifying epistemic and axiological refinement under a single formalism, we broaden the conceptual foundations of rational choice and illuminate the normative status of ethical deliberation.

Keywords: decision theory, rational choice, formal epistemology, preference change, unawareness, awareness growth, model misspecification

1. INTRODUCTION

Agents in the real world, whether human or artificial intelligence, make decisions in the face of uncertainty. One component of such decisions is *epistemic*, involving knowledge of, or beliefs about, states of the world. Another component is *axiological*, involving judgments about the value or worth of various states of affairs and the actions that can produce them. Classic value of information theorems demonstrate the importance of epistemic refinement understood as gathering information in order to change the agent's beliefs about states of the world. In particular, under certain conditions,[1] rational agents prefer acquiring non-trivial information about the state of the world. In contrast, there is little discussion of the value of axiological refinement, understood as the process of gaining greater clarity about the nature and scope of the values that determine the relative value or worth of different states of affairs. We demonstrate the importance of axiological refinement—that pausing to refine one's values, under quite general conditions, has strictly positive expected utility.

To demonstrate that axiological refinement cannot be collapsed into purely epistemic refinement, we provide results that go beyond individual decision-making contexts and show that there is strategic, not merely epistemic, value from consulting one's values. In particular, we demonstrate that in two-player zero-sum games, value refinement tends to transform interactions into ones that are positive-sum. Similarly, in Nash bargaining scenarios, value refinement tends to result in Pareto improvements for all involved parties. In these contexts, agents who refine

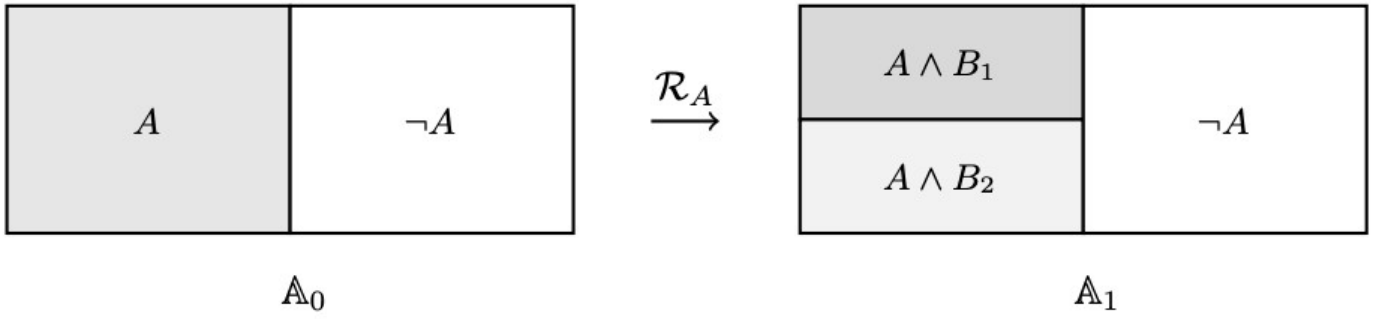
[1]Specifically, when there is act-state independence, information is costless, and agents have countably additive, proper priors (Kadane et al. 2008; Skyrms 1985; Gibbard and Harper 1978).

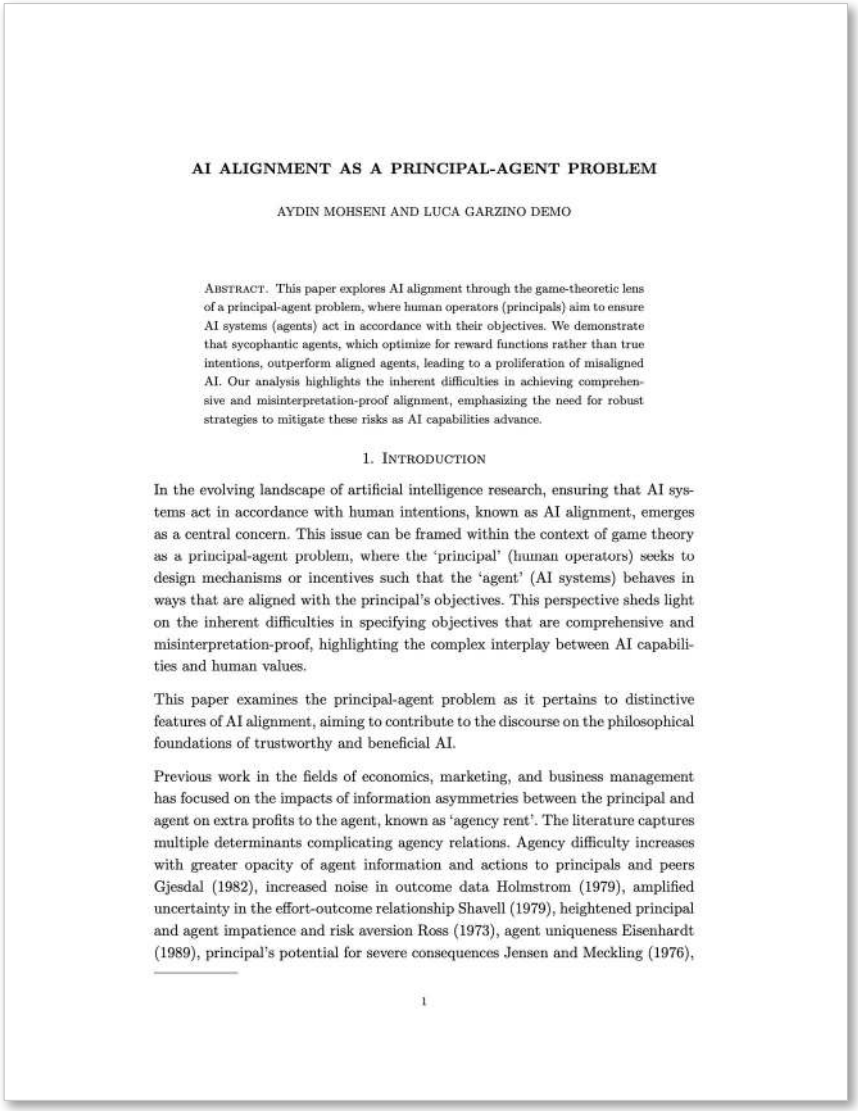
In collaboration with...



Alex John London

Carnegie Mellon University


A
$\neg A$
$\mathcal{R}_A$
$A \wedge B_1$
$A \wedge B_2$
$\neg A$
$\mathbb{A}_0$
$\mathbb{A}_1$



AI ALIGNMENT AS A PRINCIPAL-AGENT PROBLEM

AYDIN MOHSENI AND LUCA GARZINO DEMO

ABSTRACT. This paper explores AI alignment through the game-theoretic lens of a principal-agent problem, where human operators (principals) aim to ensure AI systems (agents) act in accordance with their objectives. We demonstrate that sycophantic agents, which optimize for reward functions rather than true intentions, outperform aligned agents, leading to a proliferation of misaligned AI. Our analysis highlights the inherent difficulties in achieving comprehensive and misinterpretation-proof alignment, emphasizing the need for robust strategies to mitigate these risks as AI capabilities advance.

1. INTRODUCTION

In the evolving landscape of artificial intelligence research, ensuring that AI systems act in accordance with human intentions, known as AI alignment, emerges as a central concern. This issue can be framed within the context of game theory as a principal-agent problem, where the 'principal' (human operators) seeks to design mechanisms or incentives such that the 'agent' (AI systems) behaves in ways that are aligned with the principal's objectives. This perspective sheds light on the inherent difficulties in specifying objectives that are comprehensive and misinterpretation-proof, highlighting the complex interplay between AI capabilities and human values.

This paper examines the principal-agent problem as it pertains to distinctive features of AI alignment, aiming to contribute to the discourse on the philosophical foundations of trustworthy and beneficial AI.

Previous work in the fields of economics, marketing, and business management has focused on the impacts of information asymmetries between the principal and agent on extra profits to the agent, known as 'agency rent'. The literature captures multiple determinants complicating agency relations. Agency difficulty increases with greater opacity of agent information and actions to principals and peers Gjesdal (1982), increased noise in outcome data Holmstrom (1979), amplified uncertainty in the effort-outcome relationship Shavell (1979), heightened principal and agent impatience and risk aversion Ross (1973), agent uniqueness Eisenhardt (1989), principal's potential for severe consequences Jensen and Meckling (1976),

1

In collaboration with...



- Luca Garzino Demo
  - **University of Pennsylvania**

CMU Foundations of Safe AI

(coming soon)



Vicent Conitzer

Carnegie Mellon University



Atoosa Kasirzadeh

Google Deepmind



Metascience

The aim: *a better future science*

HARKing is one the four horsemen of the replicability apocalypse [along with “publication bias, low statistical power, and *p*-hacking.”]

—Bishop, *Nature*, 2019, p. 435

HARKING: FROM MISDIAGNOSIS TO MISPRESCRIPTION

AYDIN MOHSENI

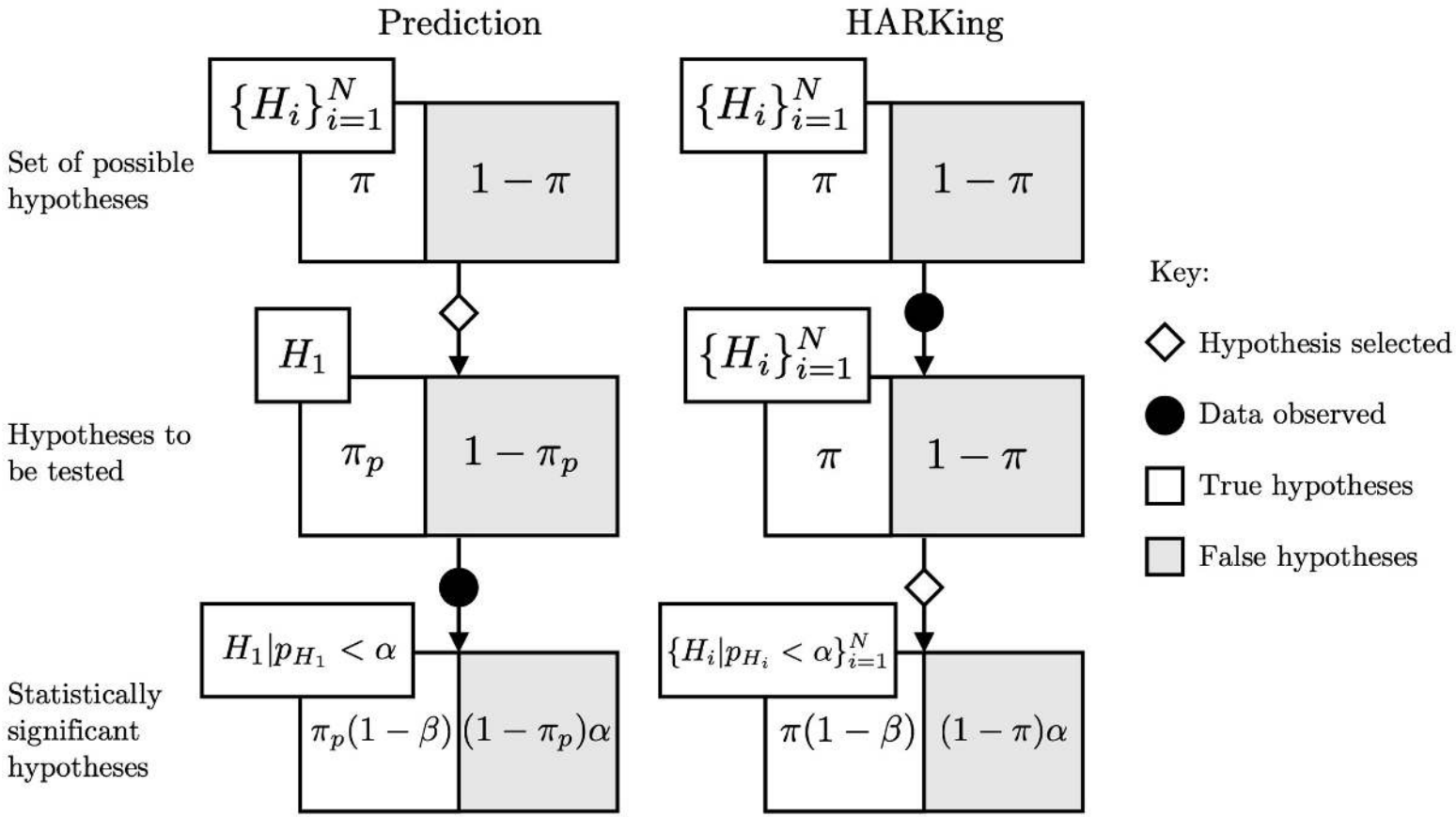
Abstract. The practice of HARKing—hypothesizing after results are known—is commonly maligned as undermining the reliability of scientific findings. There are several accounts in the literature as to why HARKing undermines the reliability of findings. We argue that none of these is right and that the correct account is a Bayesian one. HARKing can indeed decrease the reliability of scientific findings, but it can also increase it. Which effect HARKing produces depends on the difference of the prior odds of hypotheses characteristically selected ex ante and ex prost to observing data. Further, we show how misdiagnosis of HARKing can lead to misprescription in the context of the replication crisis.

1. Introduction

In a 2019 article in *Nature*, the author, psychologist Dorothy Bishop, describes HARKing as one of "the four horsemen of the reproducibility apocalypse," along with publication bias, low statistical power, and p-hacking (Bishop 2019, p. 435). The practice of HARKing—hypothesizing after results are known—is commonly maligned as undermining the reliability of scientific findings.[1] There are several accounts in the literature as to why HARKing undermines the reliability of findings. Scholars have argued that HARKing undermines frequentist guarantees of long-run error control, that it violates a broadly Popperian picture of science, and misrepresents hypotheses formulated ex post to observing the data as those formulated ex ante. We argue that none of these accounts correctly identify *why* HARKing can undermine the reliability of findings, and that the correct account is a Bayesian one.

We will show that HARKing can indeed decrease the reliability of scientific findings, but that there are conditions under which HARKing can actually increase the reliability of findings. In both cases, the effect of HARKing on the reliability of findings is determined by the difference of the prior odds of hypotheses characteristically selected ex ante and ex post to observing data. To make this precise, we employ a standard model of null hypothesis significance testing

[1] See, for example, Kerr (1998); John et al. (2012); Murphy and Aguinis (2019); and Lishner (2021). For a more nuanced account of the effects of HARKing, more in line with the one provided here, see Rubin (2017).


Prediction
HARKing
Set of possible hypotheses
Hypotheses to be tested
Statistically significant hypotheses
$\{H_i\}_{i=1}^N$
π
$1-\pi$
H_1
π_p
$1-\pi_p$
$H_1|p_{H_1} < \alpha$
$\pi_p(1-\beta)$
$(1-\pi_p)\alpha$
$\{H_i|p_{H_i} < \alpha\}_{i=1}^N$
$\pi(1-\beta)$
$(1-\pi)\alpha$
Key:
Hypothesis selected
Data observed
True hypotheses
False hypotheses

REPLICATION, REFORM, AND AN INTERVENTIONAL SIMPSON'S PARADOX

AYDIN MOHSENI

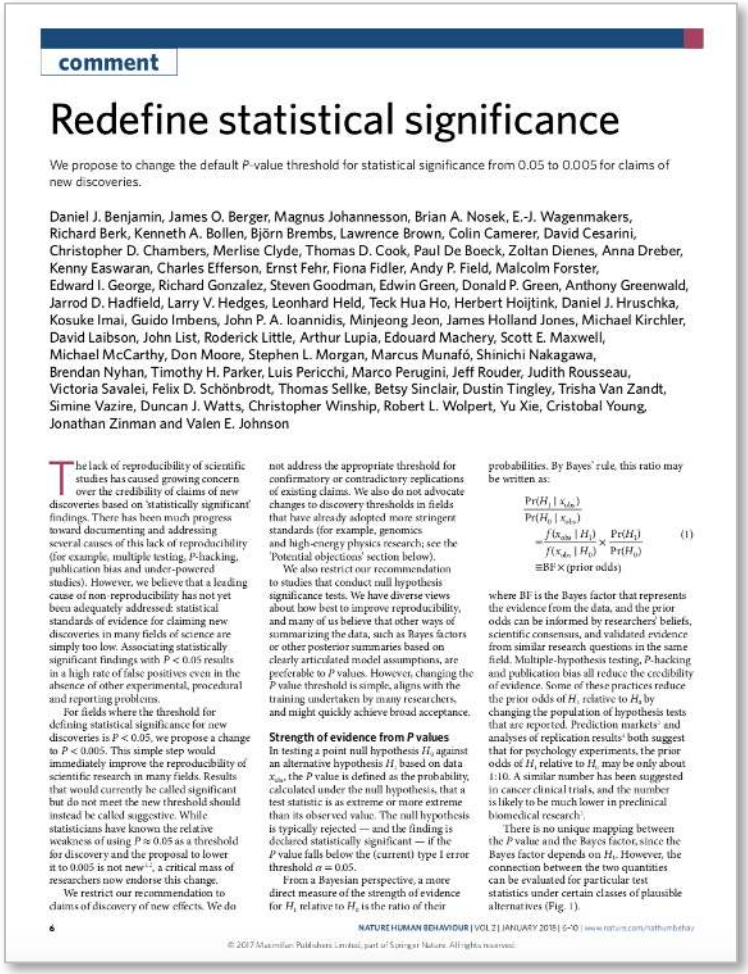
Abstract. Scientific studies vary in their methodological soundness. Interventions in evidentiary standards and research practices can differentially affect studies as a function of this soundness. The conjunction of these facts has unrecognized implications for proposed interventions in the replication crisis. We argue that we should expect these facts to obtain, and demonstrate that, when accounting for differential effects of interventions as a function of study soundness, prominent proposed interventions—lowering the significance threshold and voluntary preregistration—will produce less improvement than estimates might suggest and, in some cases, actually increase false discovery rates, sign error rates, and magnitude exaggeration ratios.

1. Introduction

A key result in social epistemology is that social rationality can diverge from individual rationality. This result, known as the Independence Thesis, was established by Mayo-Wilson et al. (2011). In their analysis of bandit problems, they show that networks of agents employing learning rules that individually fail to converge to the truth may, as a group, succeed in doing so; conversely, networks of agents using learning rules that individually converge to the truth may, collectively, fail to do so.

In this paper, we illustrate an instance of this moral in the context of the replication crisis in the social and biomedical sciences and ongoing debates about methodological reforms in scientific practice. Through an analysis of a model of null hypothesis significance testing, we show how reforms that lower every individual study's error rate can nevertheless worsen the error rate of the published literature as a whole, producing an interventional Simpson's paradox.

Our result is built on two core premises. The first premise is that scientific studies vary in their methodological soundness. Methodological soundness here is



comment

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafò, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (for example, multiple testing, *P*-hacking, publication bias and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating statistically significant findings with $P < 0.05$ results in a high rate of false positives even in the absence of other experimental, procedural and reporting problems.

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called significant but do not meet the new threshold should instead be called suggestive. While statisticians have known the relative weakness of using $P \approx 0.05$ as a threshold for discovery and the proposal to lower it to 0.005 is not new[1–3], a critical mass of researchers now endorse this change.

We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (for example, genomics and high-energy physics research; see the 'Potential objections' section below).

We also restrict our recommendation to studies that conduct null hypothesis significance tests. We have diverse views about how best to improve reproducibility, and many of us believe that other ways of summarizing the data, such as Bayes factors or other posterior summaries based on clearly articulated model assumptions, are preferable to *P* values. However, changing the *P* value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.

Strength of evidence from *P* values

In testing a point null hypothesis H_0 against an alternative hypothesis H_1 based on data x_{obs}, the *P* value is defined as the probability, calculated under the null hypothesis, that a test statistic is as extreme or more extreme than its observed value. The null hypothesis is typically rejected — and the finding is declared statistically significant — if the *P* value falls below the (current) type I error threshold $\alpha = 0.05$.

From a Bayesian perspective, a more direct measure of the strength of evidence for H_1 relative to H_0 is the ratio of their probabilities. By Bayes' rule, this ratio may be written as:

$$\frac{\Pr(H_1 \mid x_{obs})}{\Pr(H_0 \mid x_{obs})} = \frac{f(x_{obs} \mid H_1)}{f(x_{obs} \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)} \quad (1)$$
$$\equiv \text{BF} \times (\text{prior odds})$$

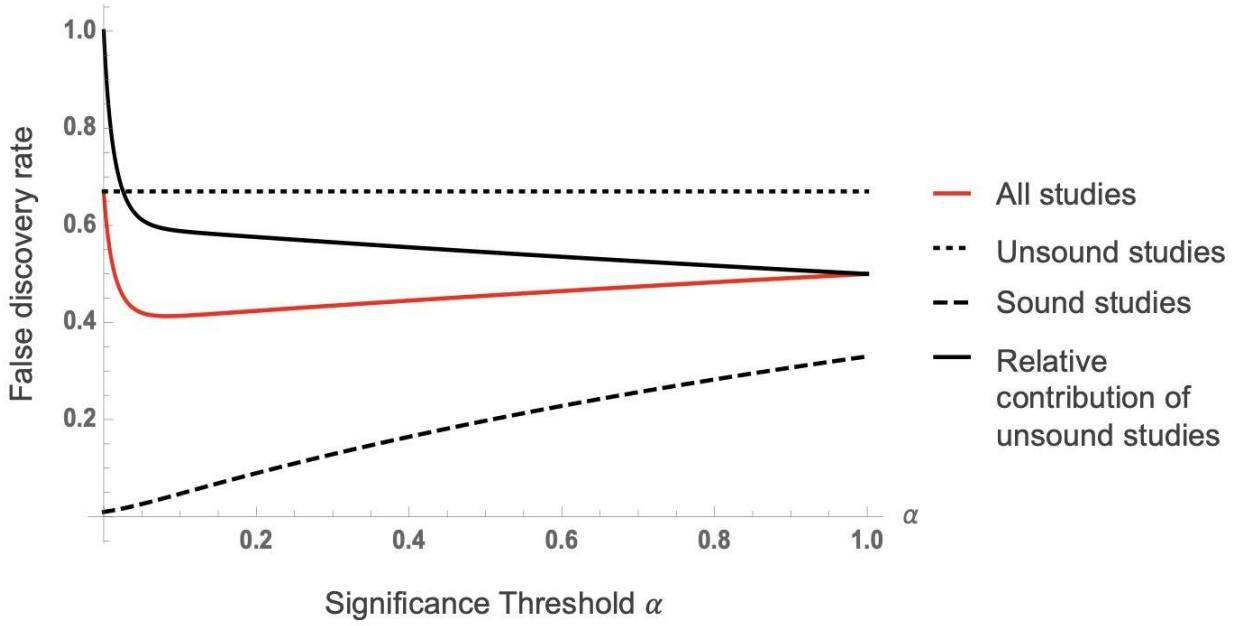
where BF is the Bayes factor that represents the evidence from the data, and the prior odds can be informed by researchers' beliefs, scientific consensus, and validated evidence from similar research questions in the same field. Multiple-hypothesis testing, *P*-hacking and publication bias all reduce the credibility of evidence. Some of these practices reduce the prior odds of H_1 relative to H_0 by changing the population of hypothesis tests that are reported. Prediction markets[4] and analyses of replication results[5] both suggest that for psychology experiments, the prior odds of H_1 relative to H_0 may be only about 1:10. A similar number has been suggested in cancer clinical trials, and the number is likely to be much lower in preclinical biomedical research[3].

There is no unique mapping between the *P* value and the Bayes factor, since the Bayes factor depends on H_1. However, the connection between the two quantities can be evaluated for particular test statistics under certain classes of plausible alternatives (Fig. 1).

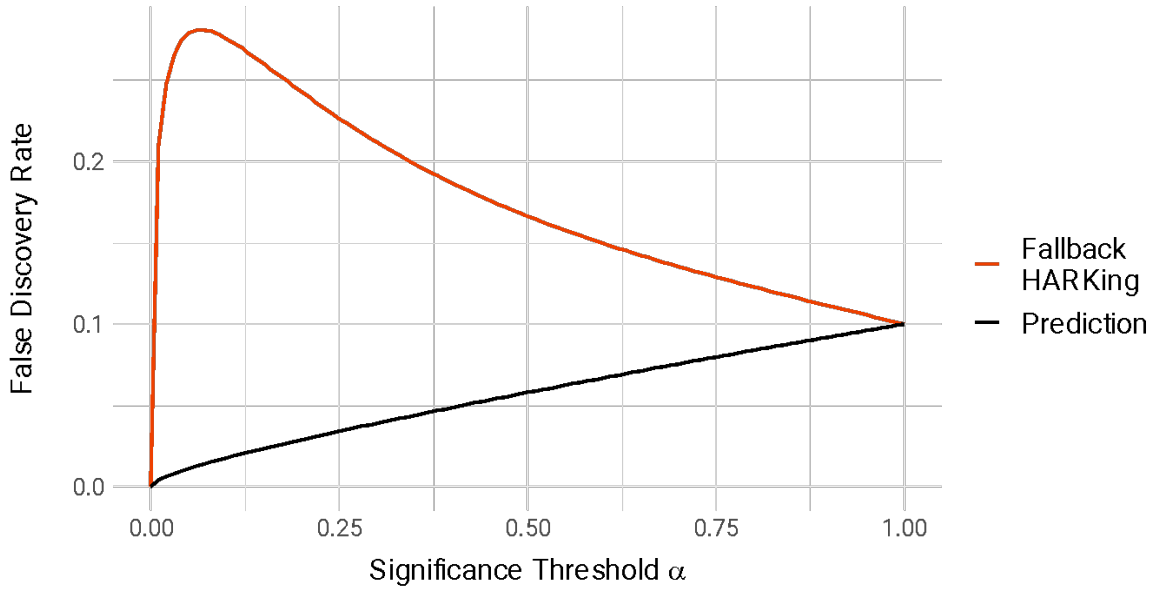
© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

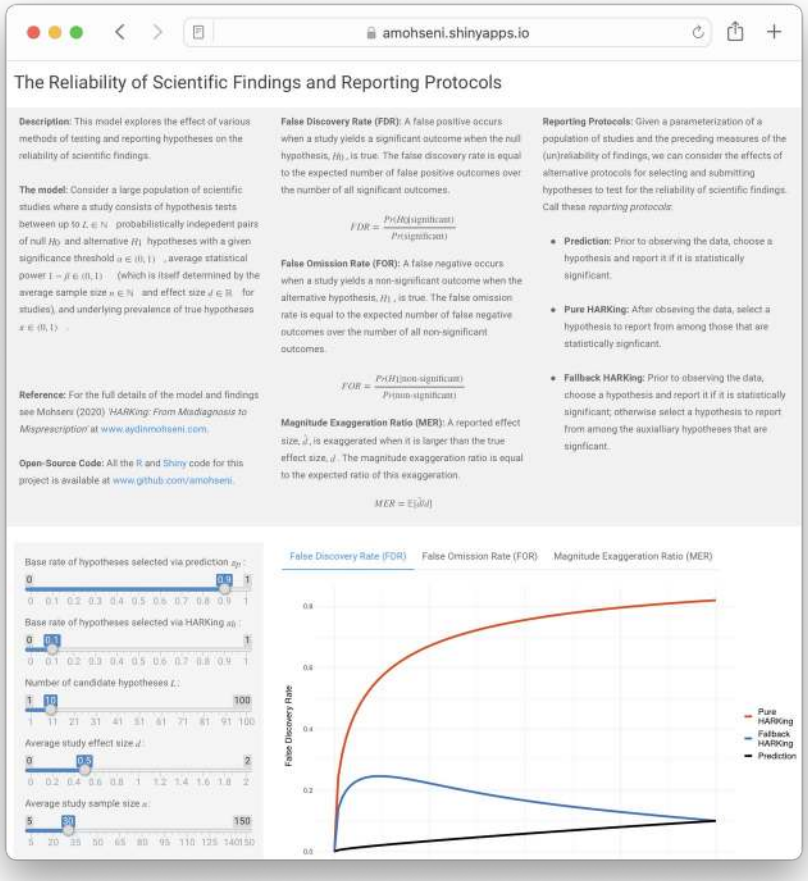
Proposal to change the default *p*-value threshold for statistical significance from 0.05 to 0.005.

+*p*-hacking


False discovery rate
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
α
Significance Threshold α
All studies
Unsound studies
Sound studies
Relative contribution of unsound studies

+HARKing


False Discovery Rate
0.2
0.1
0.0
0.00
0.25
0.50
0.75
1.00
Significance Threshold α
Fallback HARKing
Prediction



We can model the proposed reforms to help us better reason about their potential impacts.

- **False discovery rate** (currently dominates discussion)
- **False omission rate**
- **Sign error rate**
- **Absolute discovery rate**
- **Magnitude exaggeration ratio**
- **Decision-relevant informativeness**

A broader portfolio of epistemic aims

Optimal Science Funding

In collaboration with...



Kevin Zollman



Simon
DeDeo



Ísey
Mous



JOHN TEMPLETON
FOUNDATION

INSTITUTE FOR
COMPLEX SOCIAL
DYNAMICS

Grantmakers allocate ~$1T/year to fund science

We do this without a theory of what works

A funder controls three things:

- WHO (e.g., track record vs. peer review)
- WHEN (e.g., commit early vs. update later)
- SPREAD (e.g., target vs. seed everyone)

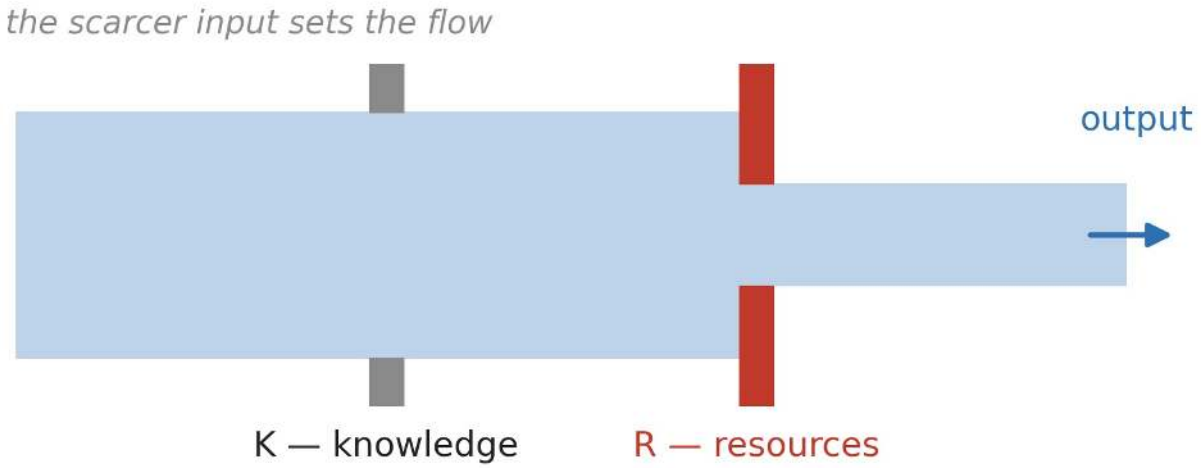
We model grantmaking as a Bayesian allocation problem.

Setup

A (human, AI) researcher is represented by two numbers:

- K = knowledge
- R = resources

Output is bottlenecked by the scarcer one

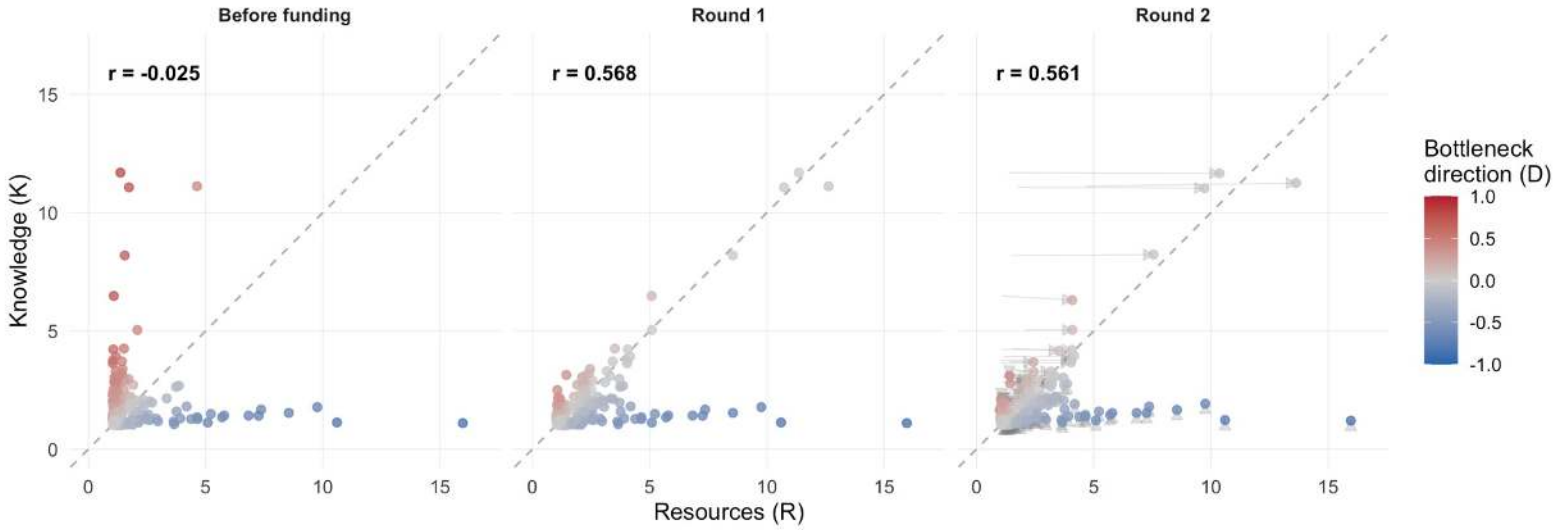


Strategy

Forward (pubs + grant)

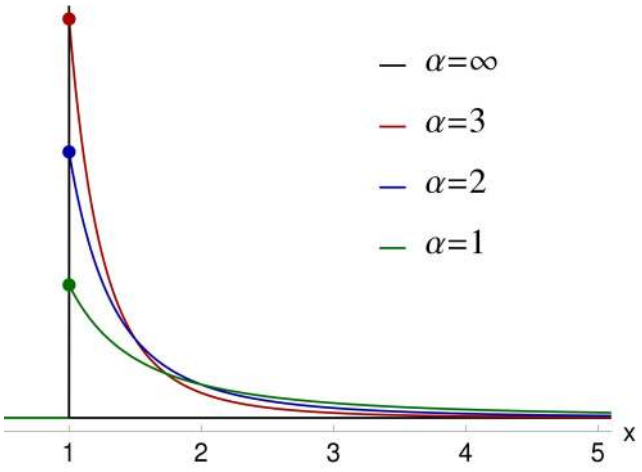
Funding effects on (K, R) distribution: Forward (pubs + grant)

B = 101/round | Arrows (right panel) show start-to-round-2 movement



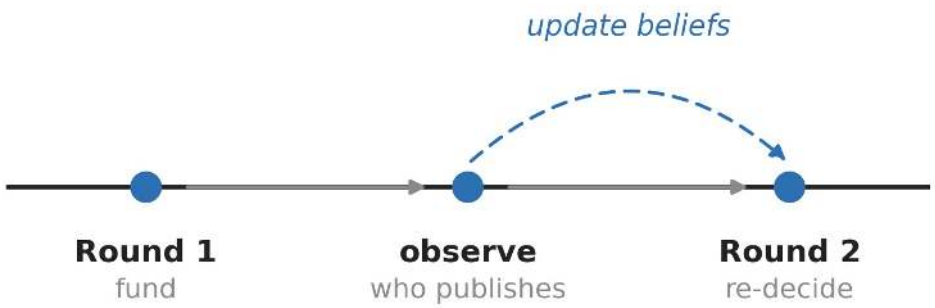
A field is a few stars and a long tail

- researchers drawn unequally
- "how unequal" is empirical ($\alpha \approx 2$)



The funder’s problem:

- a noisy signal of each researcher
- split a budget over a few rounds
- see who produces what, then re-decide

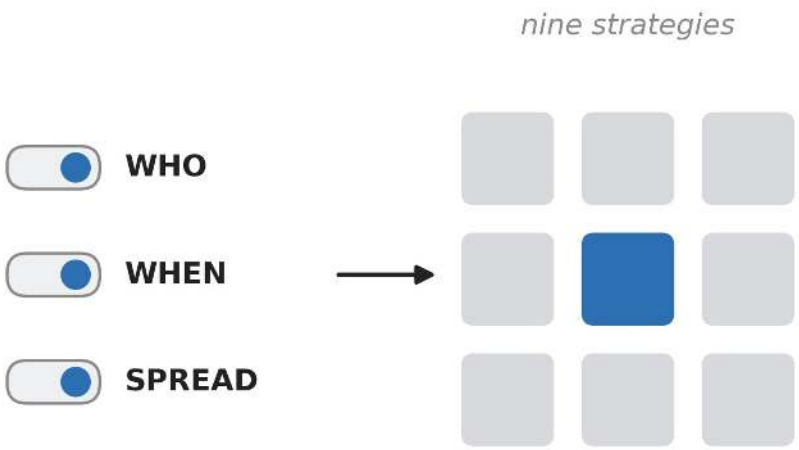


What the optimal dollar balances:

- return now
- build them up
- return later
- learn by funding
- learn by watching

Three knobs & nine strategies:

- Each strategy = a knob setting
- Comparing them is the experiment



Results

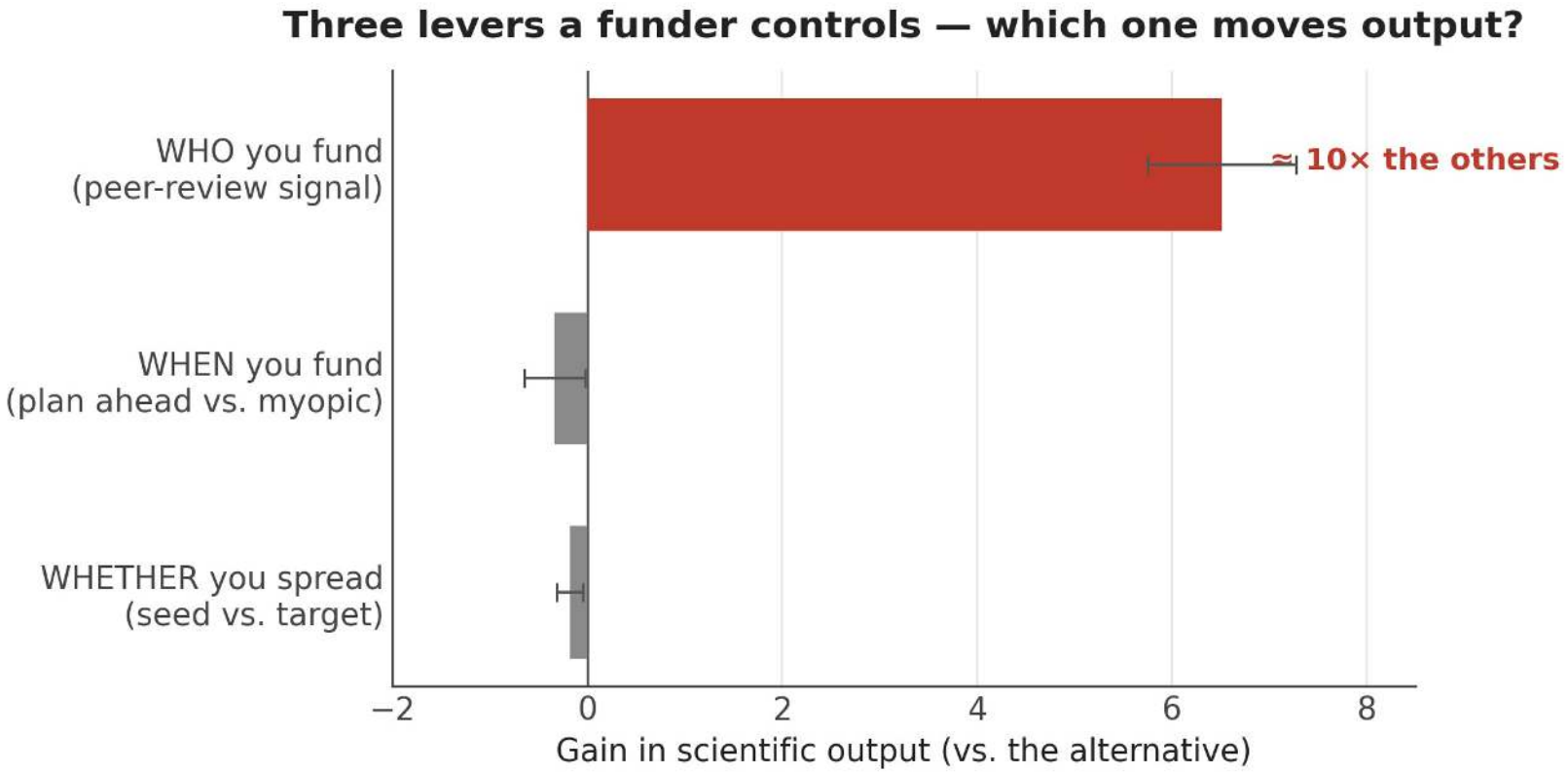
These are some very preliminary results

Let’s make some predictions...

- six questions
- predict before each reveal

Which lever buys the most science?

- WHO (e.g., track record vs. peer review)
- WHEN (e.g., commit early vs. update later)
- SPREAD (e.g., target vs. seed everyone)

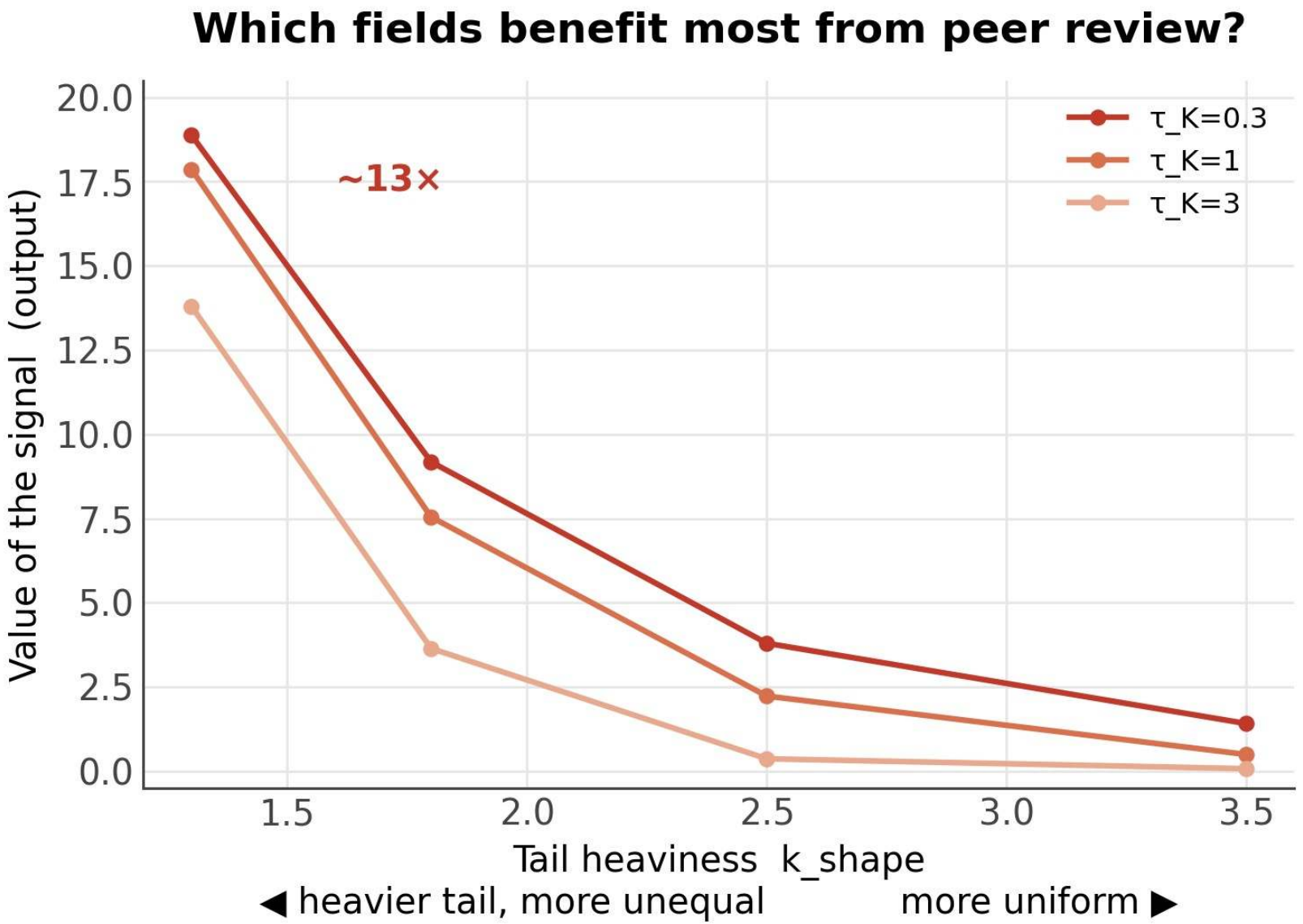


WHO ≈ 10× WHEN or SPREAD

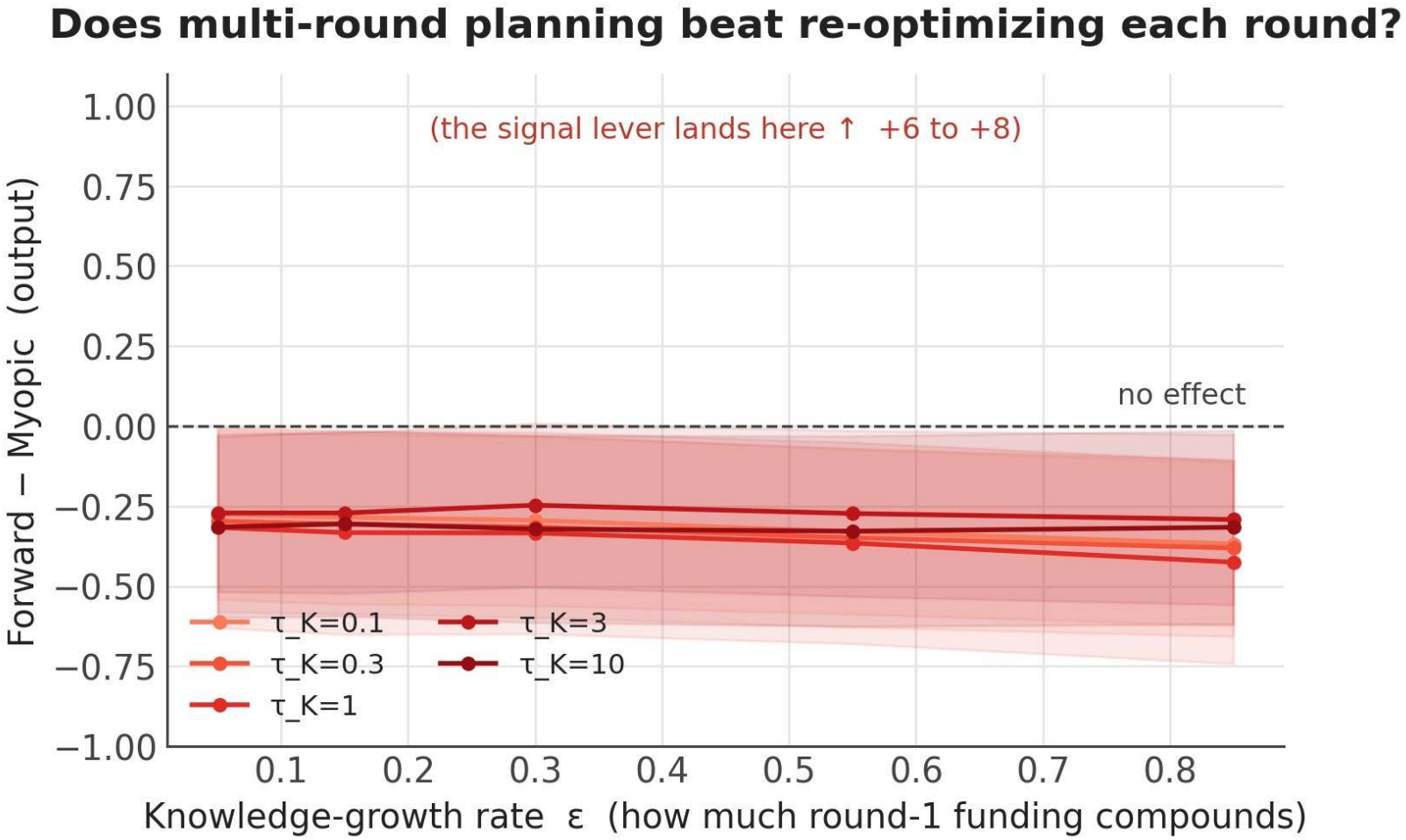
How sharp does peer review need to be?


How good does peer review need to be?
Value of the signal (output)
9
8
7
6
5
4
3
2
1
0
further precision
barely helps
10^{-1}
10^{0}
10^{1}
Signal noise τ_K
◀ sharper review
noisier review ▶

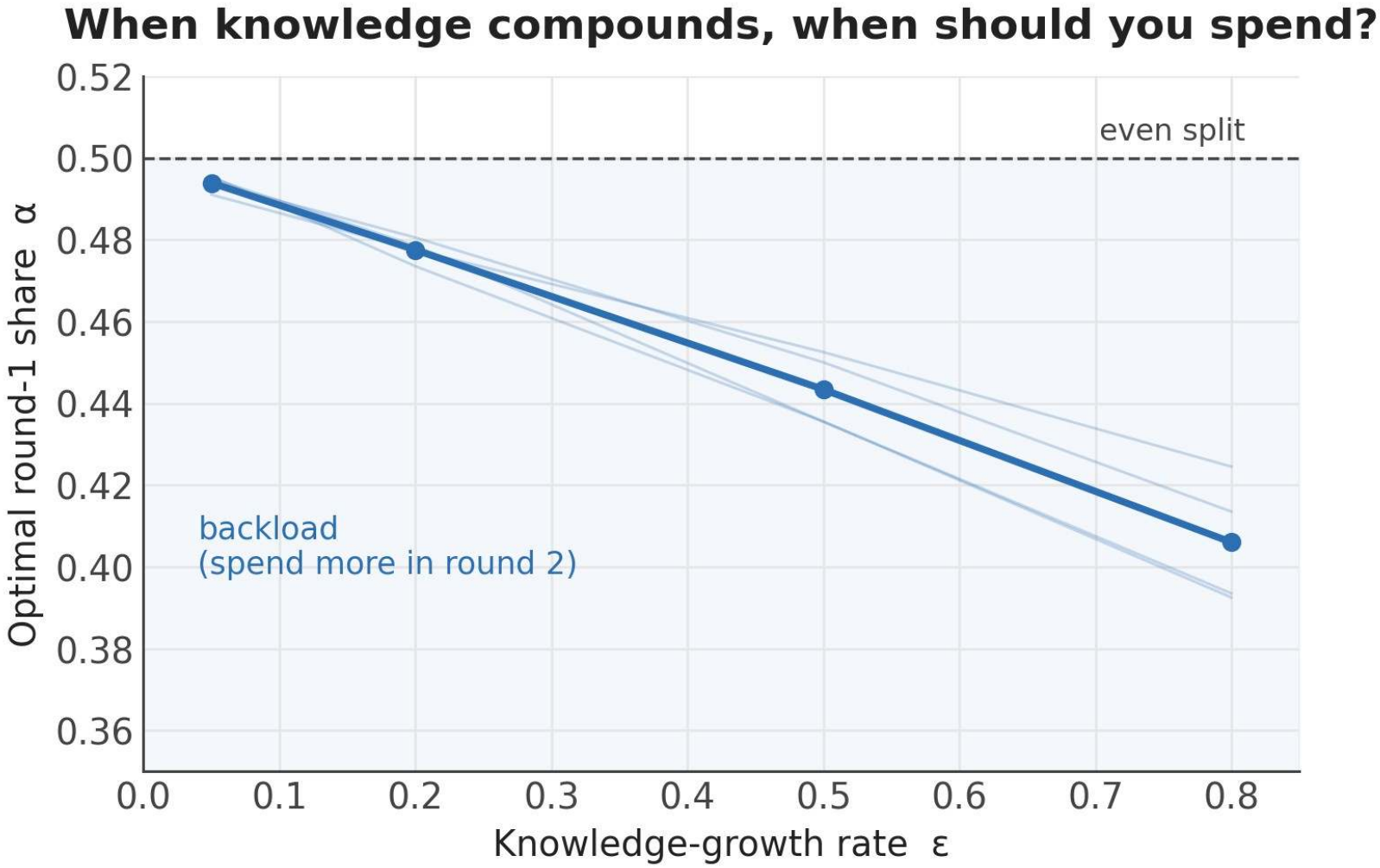
Does review pay off equally everywhere?


Which fields benefit most from peer review?
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Value of the signal (output)
~13×
τ_K=0.3
τ_K=1
τ_K=3
1.5
2.0
2.5
3.0
3.5
Tail heaviness k_shape
◀ heavier tail, more unequal
more uniform ▶

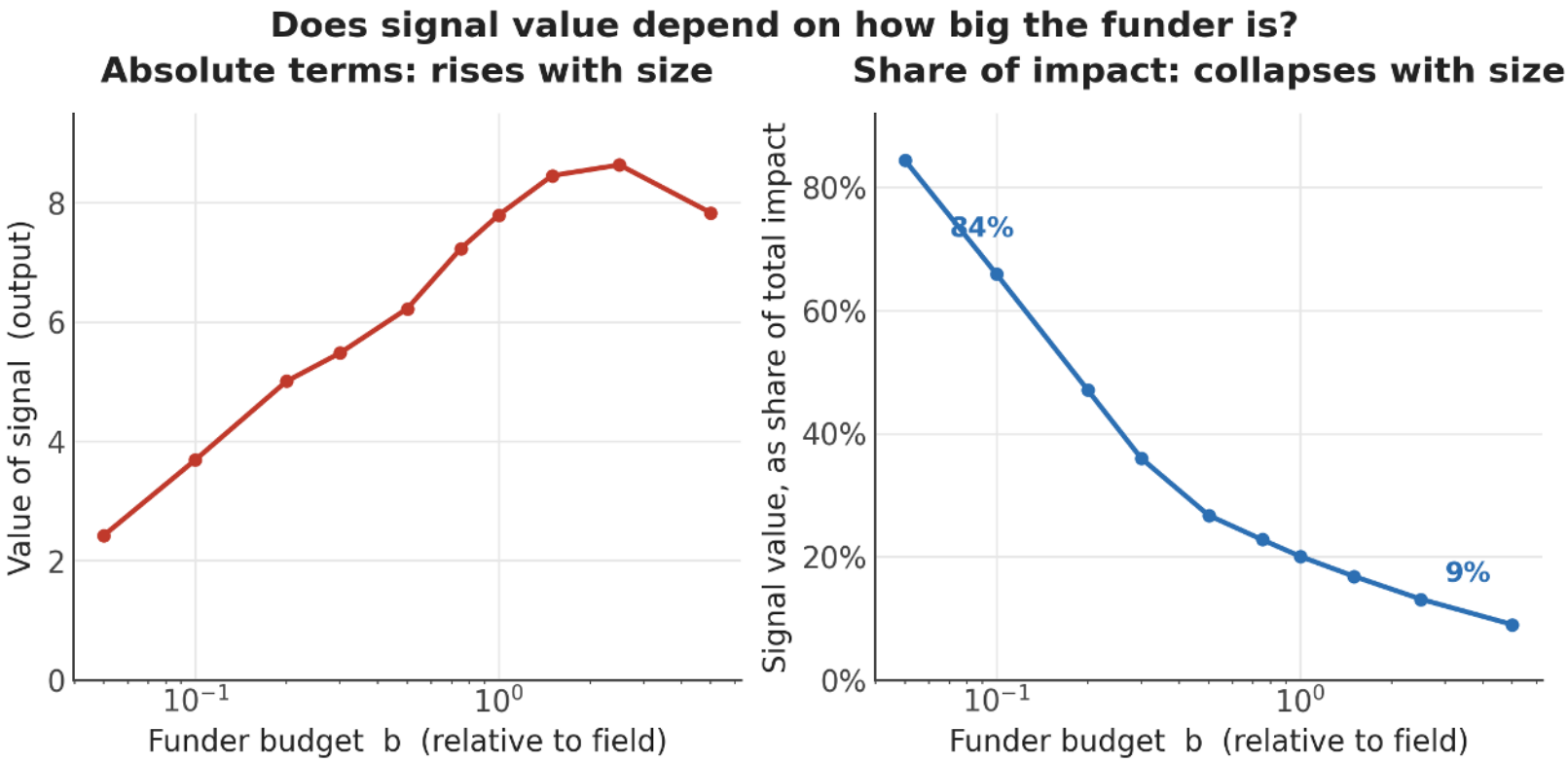
Does planning ahead beat re-deciding each round?


Does multi-round planning beat re-optimizing each round?
(the signal lever lands here ↑ +6 to +8)
no effect
Forward − Myopic (output)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
τ_K=0.1
τ_K=0.3
τ_K=1
τ_K=3
τ_K=10
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Knowledge-growth rate ε (how much round-1 funding compounds)

Invest early to compound—right?


When knowledge compounds, when should you spend?
0.52
0.50
0.48
0.46
0.44
0.42
0.40
0.38
0.36
Optimal round-1 share α
even split
backload
(spend more in round 2)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Knowledge-growth rate ε

Does a small foundation face NIH's problem?


Does signal value depend on how big the funder is?
Absolute terms: rises with size
Share of impact: collapses with size
Value of signal (output)
0
2
4
6
8
10^{-1}
10^{0}
Funder budget b (relative to field)
Signal value, as share of total impact
0%
20%
40%
60%
80%
84%
9%
10^{-1}
10^{0}
Funder budget b (relative to field)

A caveat...

Discussion

Two bets we haven’t run

F · does upstream funding (NIH/NSF) shrink what we add?

G · which strategy wins where?

Same machine, different substrate

— money → human scientists

— compute → AI scientists

Help me: how do we quantify "quality"?

— output here = papers

— breakthroughs? training? replication?

— could a market price the gap?

How to model signals?

— mine = clean Gaussian noise

— real review = biased, strategic, committee-filtered

— does "signal dominates" survive?

Does this transfer to AI?

— right production function for automated research?

— does fixed-population break (effort, entry)?

— which single result would you bet against?

Discussion