RESEARCH ON RESEARCH INSTITUTE

Funder experimentation with AI
EViR 22 January 2026

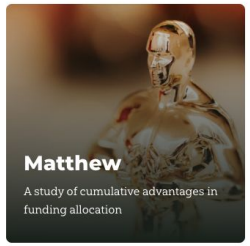
Tom Stafford, Senior Research Fellow

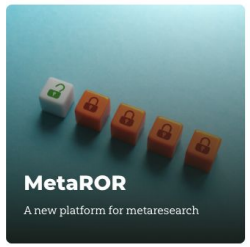
t.stafford@researchonresearch.org

tomstafford.github.io

RESEARCH ON RESEARCH INSTITUTE

12 codesigned projects





19 Core Partners

Australian Government
Australian Research Council

Michael Smith Health Research BC

NRF National Research Foundation

Fundació "la Caixa"

Swiss National Science Foundation

CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

novo nordisk foundation
Benefitting people and society

NWO

CWTS Meaningful metrics

wellcome

DIGITAL science

VolkswagenStiftung

GORDON AND BETTY MOORE FOUNDATION

King Baudouin Foundation
Working together for a better society

Social Sciences and Humanities Research Council of Canada

Canada

FWF Austrian Science Fund

Luxembourg National Research Fund

UKRI Research England

UCL

The Research on Research Institute Impact Report (2023-2025)

https://researchonresearch.org/roris-impact-what-weve-achieved-and-where-were-going/

AFIRE: Accelerator for Funder Experimentation

Forum — **Sharing work by funders, for funders**

Capacity building — Sprints on **AI/ML in reviewer selection**

Experiments — **Distributed Peer Review, Partial Randomisation, Desk Rejection, and more!**

RoRI RESEARCH ON RESEARCH INSTITUTE

IGL Innovation Growth Lab by nesta

These slides:
bit.ly/tomstafford

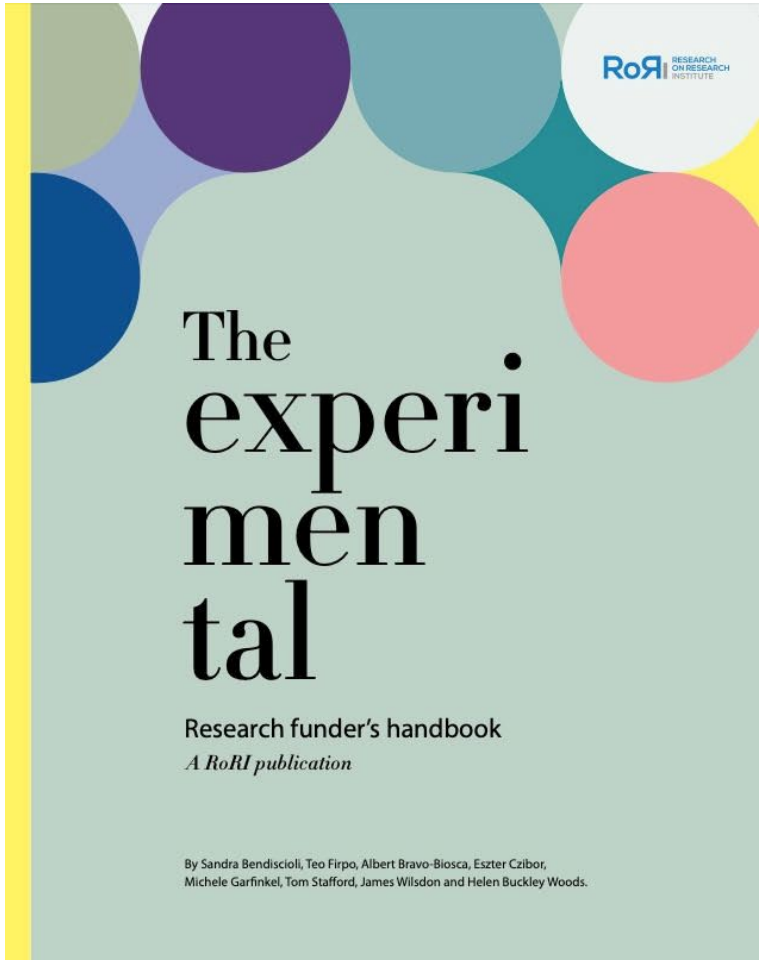
Our definition of experiment

Principled: a research design that allows inference about what causes what (before/after, shadow experiments, true experiment/RCT)

Planned: primary outcome measure and analysis plan declared in advance

Public: a commitment to sharing the results regardless of outcome

We can plan/run experiments



AFIRE:

t.stafford@researchonresearch.org

josie.coburn@ucl.ac.uk

The experimental research funder's handbook (Revised edition, June 2022, ISBN 978-1-7397102-0-0). https://doi.org/10.6084/m9.figshare.19459328.v2

These slides:

bit.ly/tomstafford

RoRI's GRAIL project

GRAIL = Getting responsible about AI and machine learning (ML) in research funding and evaluation

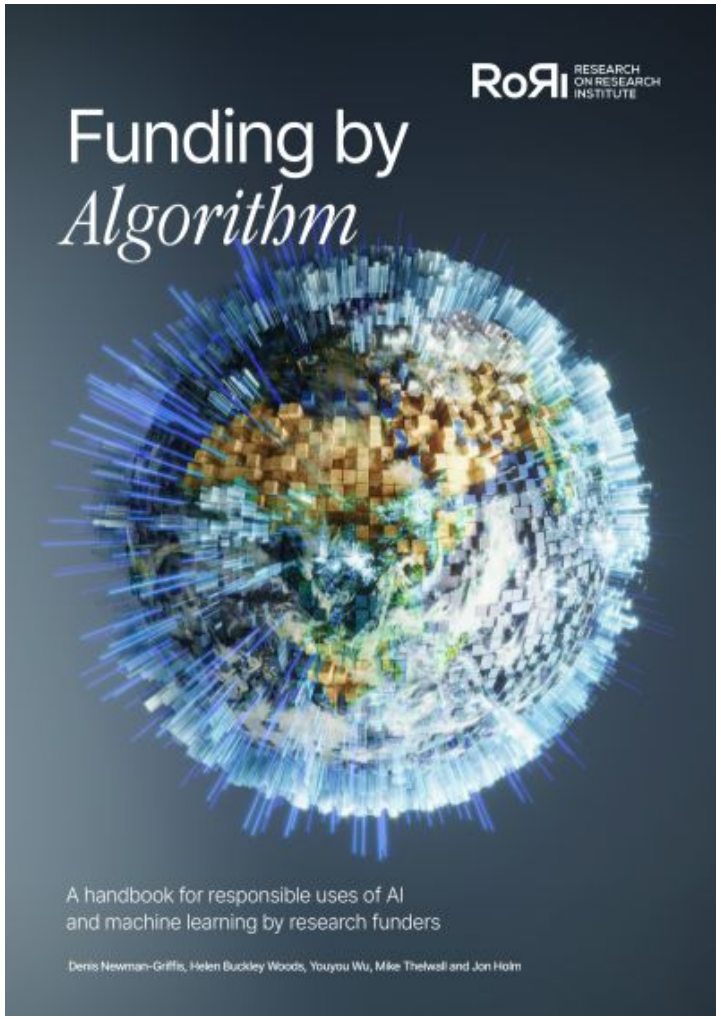
A RoRI project running from 2023 to 2025

Goal: Understand how funders use AI/ML and build shared practices

Who:

- 35 delegates from 13 research funders
- 4 researchers

Funding By Algorithm



A handbook for responsible uses of AI and machine learning by research funders

[link]

Sprint v1.0 - 20 x N America and European funders

AI in reviewer matching	3.8
AI in peer reviewing	3.5
AI for prioritising funding applications	3.3
AI in research assessment exercises	3.2
AI for applicant self-assessment	3.2
AI for navigating funding resources	3.0
AI in strategic planning	2.8

Three case studies explore real-world approaches and trade-offs: https://tinyurl.com/y8a6vfdn

Swiss National Science Foundation

"la Caixa" Foundation

NWO

AI reviewer matching@Metascience2025

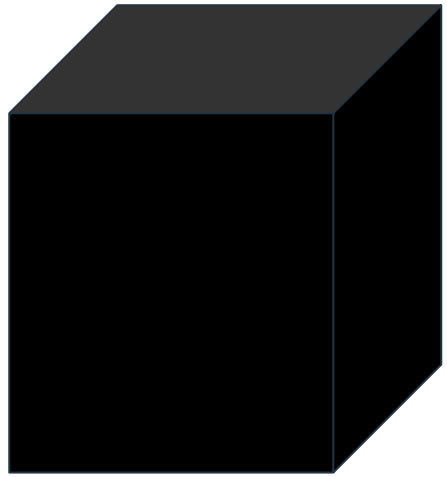
Research Questions

1. Can language models help match proposals to reviewers? → Maybe - evidence for meaningful improvements beyond human matching
2. Is it feasible for something like a conference to adopt/adapt this technology? → Definitely yes
3. Can it be done securely/privacy respecting? → Definitely yes

Why AI specifically requires experiments

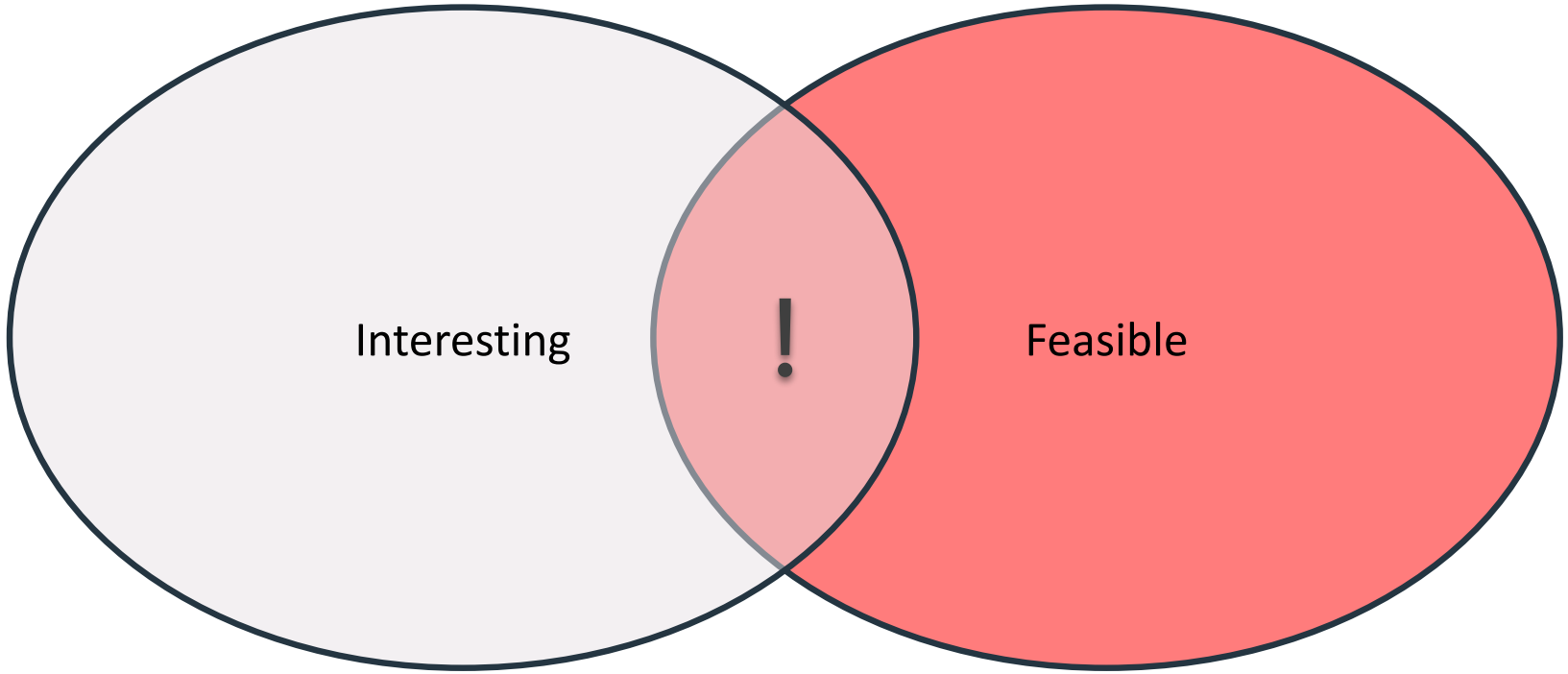
AI is novel & changing - our intuitions are poorly calibrated

Specific areas of uncertainty
- Fluency
- Stochasticity
- Hallucinations
- Inscrutability
- Bias



Thinking about experiments

"The Art of the soluble"



Good outcome measures

A part of the fresco “Triumph of Galatea,” created by Raphael around 1512 for the Villa Farnesina in Rome. Art Images via Getty Images

Assays and microscopes

Yes/No

but is it the right question

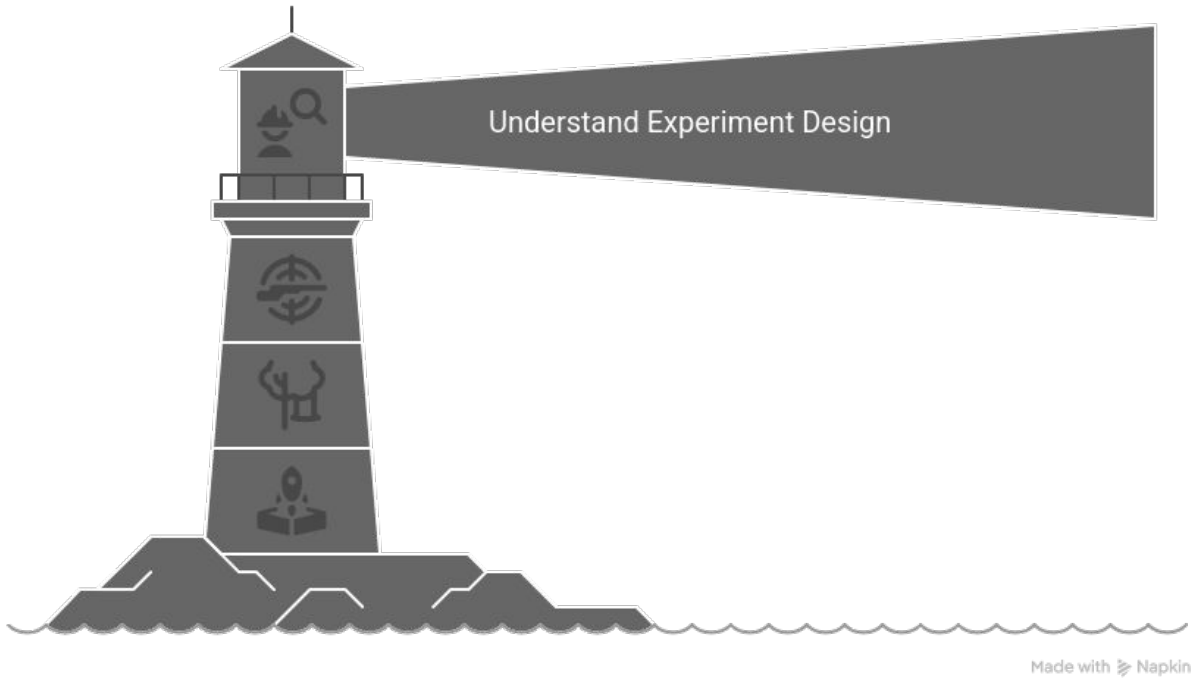
Close view

but what are you looking for?

Image: CC Wikimedia

Just designing experiments is valuable

Becoming experimental



Get in touch!

researchonresearch.org
@RoRInstitute

t.stafford@researchonresearch.org

RoRI RESEARCH ON RESEARCH INSTITUTE

END
(reserve slides follow)

Other AFIRE experiments

Partial Randomisation Trials Catalogue

Funder	Dates
Health Research Council of New Zealand	2013-
VolkswagenStiftung	2017-2020
Austrian Science Fund (FWF)	2019-
Swiss National Science Foundation (SNSF)	2018-
Novo Nordisk Fonden	2022-2025
British Academy	2022-2025
UKRI / NERC	2022-
Wellcome	2023-
Nesta	2019-2020
University of Leeds	2023
UMC Utrecht/Ministry of OCW	2023

bit.ly/PRtrials

Desk Rejection Shadow Experiment

Can agency staff predict those proposals with the least likelihood of success?

- a shadow experiment, not an intervention
- supports optimal use of external review
- UKRI leading participation
- recruiting schemes which will complete by end of 2026

Enquiries:
Josie Coburn,
Research Fellow in Metascience,
Research on Research Institute
josie.coburn@ucl.ac.uk

Evaluating Distributed Peer Review at the Volkswagen Foundation

Anna Butters, Melanie Benson Marshall, Tom Stafford & Stephen Pinfield (Research on Research Institute and University of Sheffield);
Hanna Denecke, Alexander Bondarenko, Barbara Neubauer, Robert Nuske & Pierre Schwidlinski (Volkswagen Foundation)

RoRI RESEARCH ON RESEARCH INSTITUTE

VolkswagenStiftung

More on GRAIL & the GRAIL case studies

GRAIL activities

13 virtual workshops

Presentations / case studies

Q&A

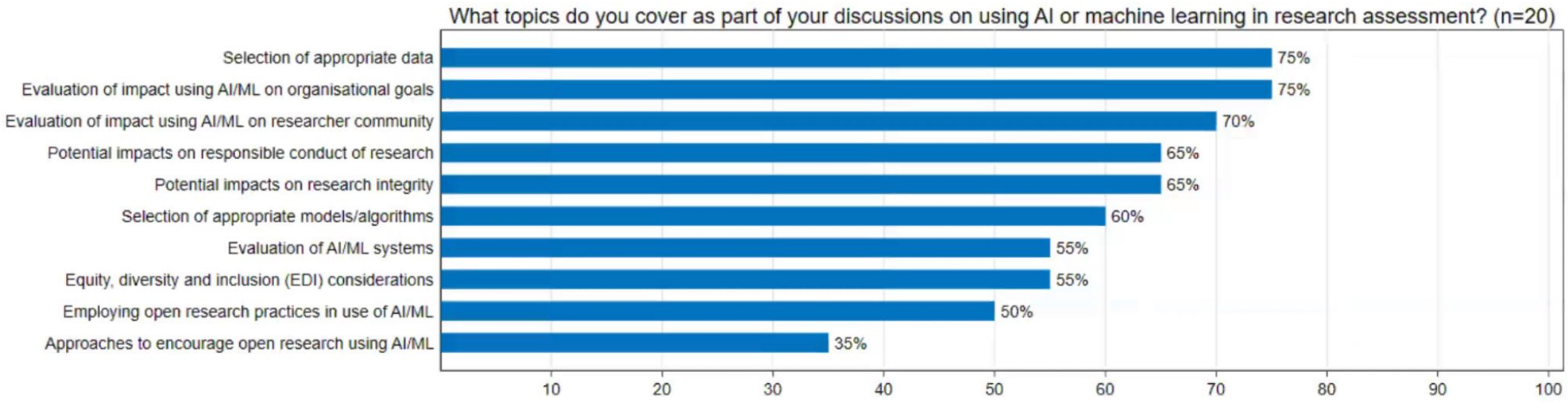
Group discussions

	Topic
7	Guidelines for the use of generative AI in research funding processes
8	Responsible AI principles for research funders
9	Human in the Loop
10	Collaboration and reuse: tools, data, and knowledge structures
11	Competencies and collaboration on AI/ML applications
12	Impact assessment, documentation and reporting, and transparency and reliability

GRAIL outputs

A curated library of AI/ML use cases

Survey: experiences with AI/ML



GRAIL insights: a wide range of AI use cases

Peer review

Handling applications

Tracking and evaluating research outputs

Assisting applicants

GRAIL insights: AI use cases

Peer review

- Matching proposals to reviewers [Handbook case study 1]
- Evaluating the quality of reviewers' comments
- Summarising and integrating reviewer comments

GRAIL insights: AI use cases

Handling applications

- Screening and prioritising applications [Handbook case study 2]
- Automatic tagging of proposals for topic/theme/SDGs
- Summarising / translating proposals
- Detecting duplicate / similar applications
- Verify eligibility

GRAIL insights: AI use cases

Tracking and evaluating research outputs

- Linking funded grant and research publications [Handbook case study 3]
- Linking funded grant and research impacts (patent/policy/media) [Handbook case study 4]
- Quality assessment of research publications [Handbook case study 5]

GRAIL insights: AI use cases

Assistance for applicants

- Chatbot for understanding funding calls
- Proposal writing assistant
- Proposal review assistant
- Self-evaluation of proposal

GRAIL insights: evaluation and experimentation

GRAIL has seen some excellent examples of experiments (a few among the participants)

A few suggested areas for evaluation in the handbook

- AI in reviewer matching
- AI in producing peer review reports
- AI for prioritising funding applications
- AI in research assessment exercises
- AI for applicant self-assessment
- AI for navigating funding resources
- AI in strategic planning

GRAIL insights

More systematic experiments with AI/ML tools is needed

Small, focused experiments will help us move beyond anecdotes

Potential for multi-funder collaboration on experiments

Swiss National Science Foundation

Implementation

Since 2020 in the Life sciences (1000 proposals/year)

Fixed pool of reviewers

Full text of proposals

In house data team

Considerations

Human in the loop (all matches reviewed)

Explainability (not generative)

Transparency (open, not commercial models)

Privacy (run locally)

Swiss National Science Foundation

Evaluation:

Informal: time saving

Formal: overlap with human matching (“ in more than 8 out of 10 proposals, there was at least one reviewer suggested by the AI-assisted reviewer matching that was also manually assigned by the scientific officers.”)

Learnings

Demonstrate feasibility & efficiency *as a supporting tool*

Heterogeneity in performance

Integration with organisation requires as much attention/resource as technical implementation

Okasa, G., & Jorstad, A. (2024). The Value of Pre-training for Scientific Text Similarity: Evidence from Matching Grant Proposals to Reviewers. In Proceedings of the 9th edition of the Swiss Text Analytics Conference (pp. 89-101).

"la Caixa" Foundation

Implementation

Last 8 cycles of CaixaResearch Health call, also now Innovation call

Fixed reviewer pool

Co-designed/developed with external provider

Keywords (MeSH) from proposals sent to external provider, reviewer keywords obtained from pubmed, matches returned

Considerations

Time/workload saving,

Accuracy/fairness

Population balance (e.g. gender assignment)

"la Caixa" Foundation

Evaluation

10% of matches reviewed by human

Audit of % of reviewers who declare a lack of expertise on assigned proposals (stable at 20% per year)

Learnings

Successful for improving not replacing human intervention

Other uses of AI: detecting proposals with low probability of being funded, reviewing evaluators' comments, summarizing selected proposals to produce abstracts for the general public and summarizing evaluators' comments

Implementation

Tool available organisation wide

prophy.ai/ (previously Elsevier tool)

Proposals & applicant details uploaded, reviewer suggestions come back (no pool), used to augment existing reviewer search methods

Considerations

Prophy driven by OA sources

Motivations: Speed, widen reviewer pool, improve match quality

Evaluation

No formal evaluation - difficulties is assessing value for staff and in assessing quality of reviews resulting

Learnings

tool won't tell you reviewer quality, or timeliness

perceived to be successful (a useful tool, better than keyword only search)

external party easier than in house development

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e18, 1-25
DOI: 10.22163/fteval.2025.710

INSIDE THE FUNDING PROCESS: USING GENERATIVE AI TO ASSESS REVIEWERS' CRITERIA PRIORITISATION IN MULTI-STAGE APPLICATION ASSESSMENTS

PETER KOLARZ AND DIOGO MACHADO
DOI: 10.22163/FTEVAL.2025.710

FWF Austrian Science Fund

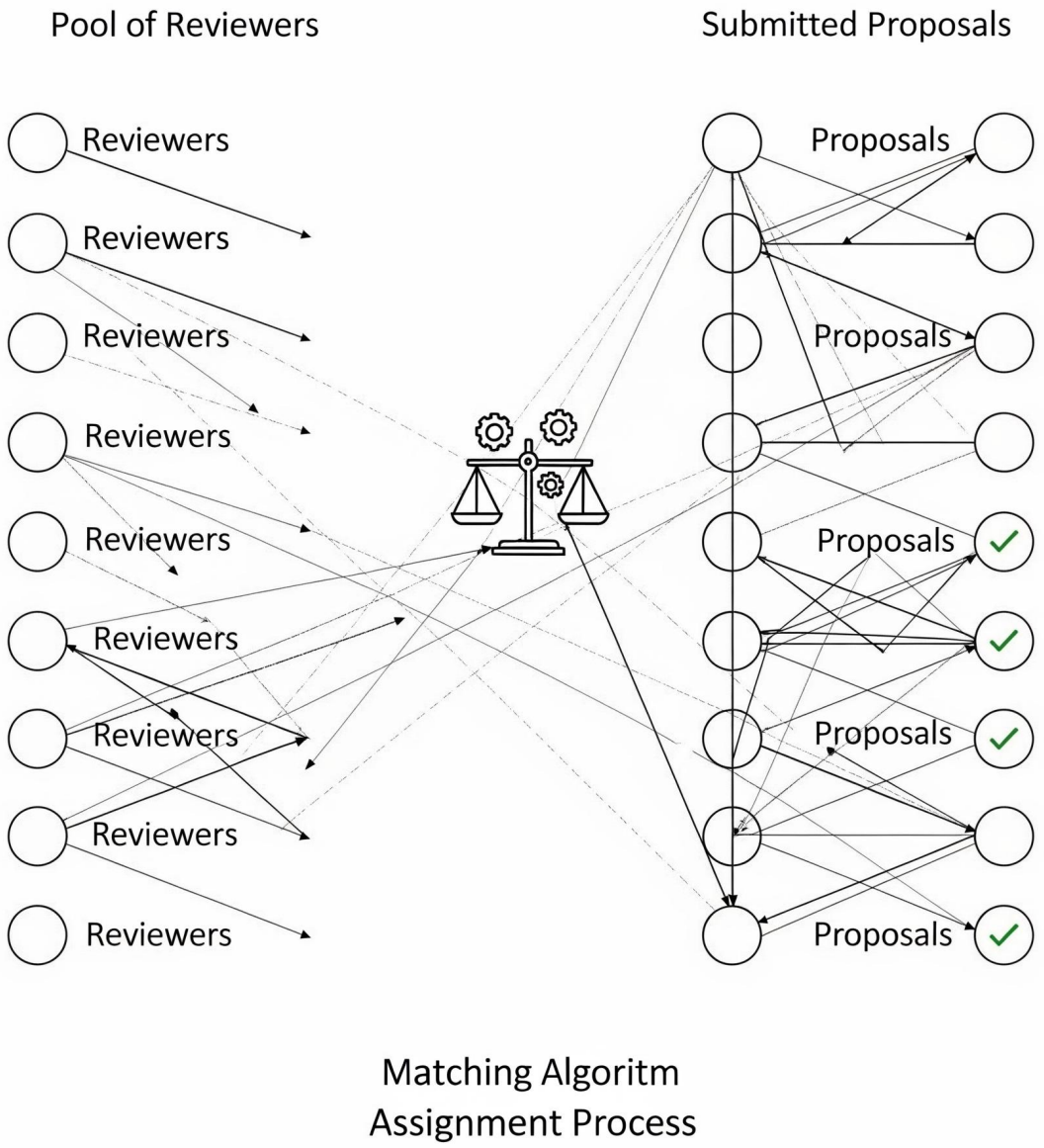
RoRI RESEARCH ON RESEARCH INSTITUTE

The Metascience 2025 conference experiment

RoRI RESEARCH ON RESEARCH INSTITUTE

Can AI be used for better matching of proposals to reviewers? Feasibility and formal evaluation with the Metascience 2025 conference

Josie Coburn and Tom Stafford 2025-11-26



Finding (enough, good) reviewers is a conceptual and practical problem

- conceptual: what makes a reviewer good?
- practical: how do you get a reviewer to agree?

Reviewer-proposal matching identified by GRAIL as a key area for possible experiments

Many funders already exploring this

Algorithms need validation!

Meta-metascience

AFIRE Commitment: Observation is not enough - we have to try things!

- demonstrate feasibility
- opportunity for better causal inference

Metascience 2025 conference, London

- a chance to show we'll take our own medicine
- appropriate domain for demonstrating feasibility
- added value: validate by collecting reviewer self-perception of suitability

The “shadow” experiment

Consent from those submitting and reviewers

All analyses done after final programme decisions

All analyses local - no data left the conference

441 submissions: Title, Abstracts

25 reviewers: assigned to submissions via keywords

1,323 reviews

- for each we have a match scores & a reviewer suitability judgement
- (each proposal seen by 3 reviewers)

Research Questions

1. Can language models help match proposals to reviewers?

2. Is it feasible for something like a conference to adopt/adapt this technology?

3. Can it be done securely/privacy respecting?

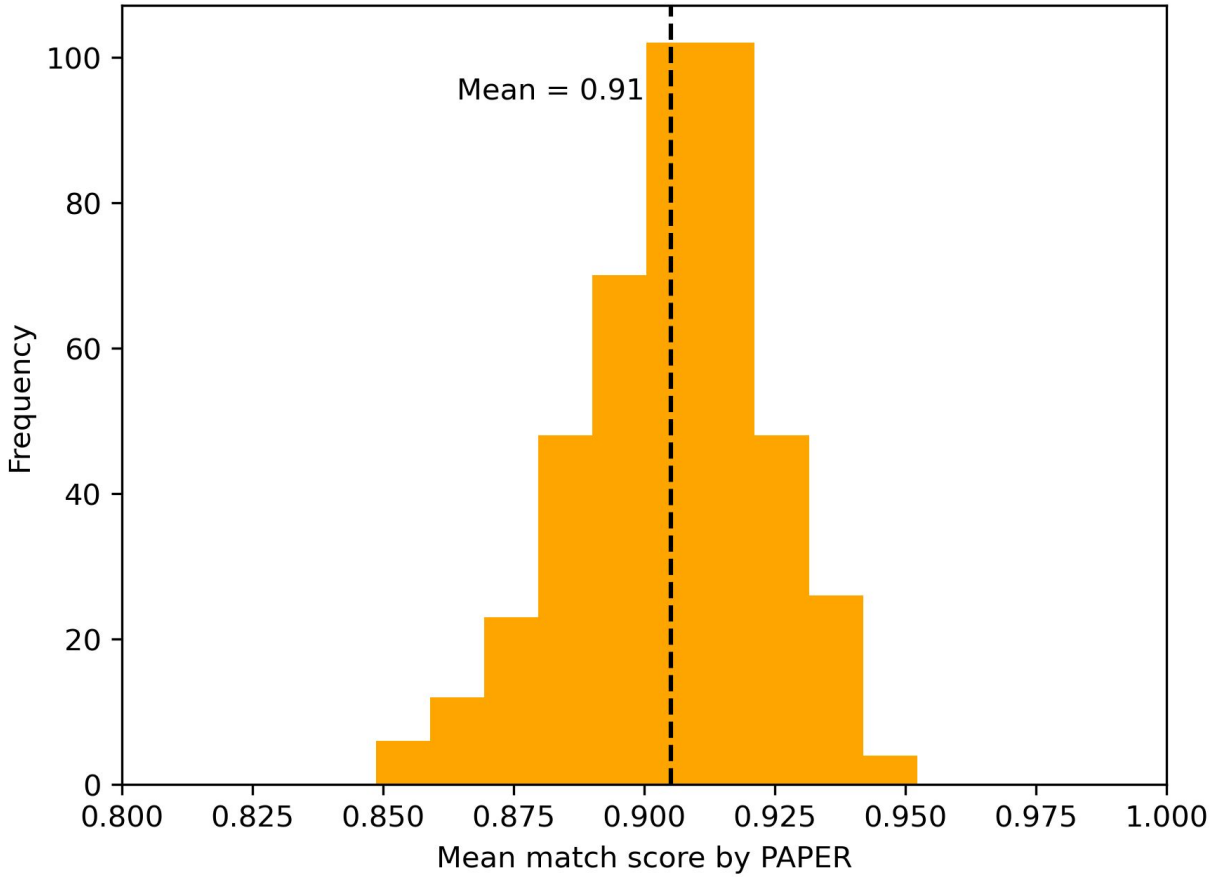
Matching - via embedding

Reviewer keywords & proposal title+abstract -> embedding space

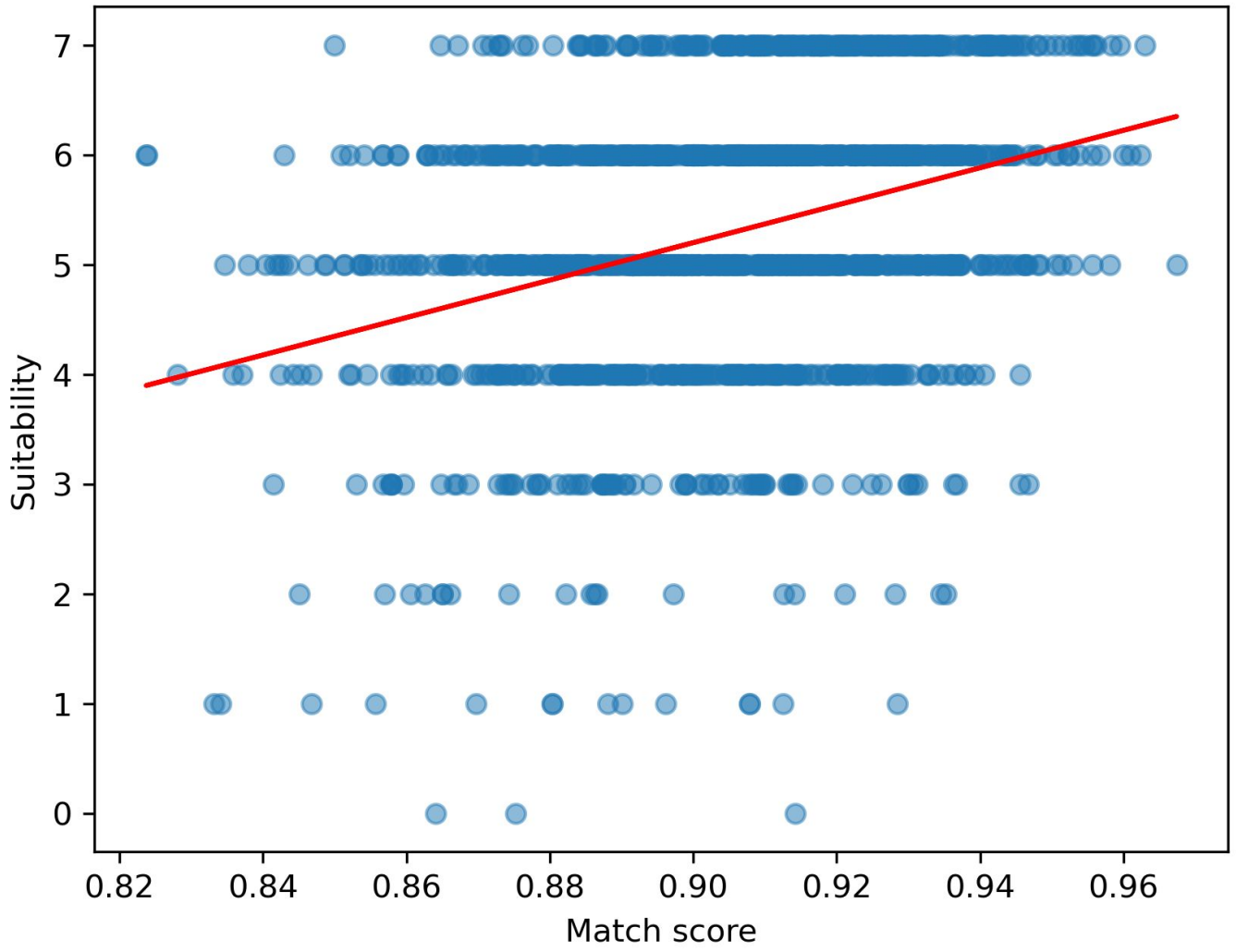
Code from SNSF: https://github.com/snsf-data/snsf-grant-similarity

- thanks to Gabriel Okasa and the SNSF data team!

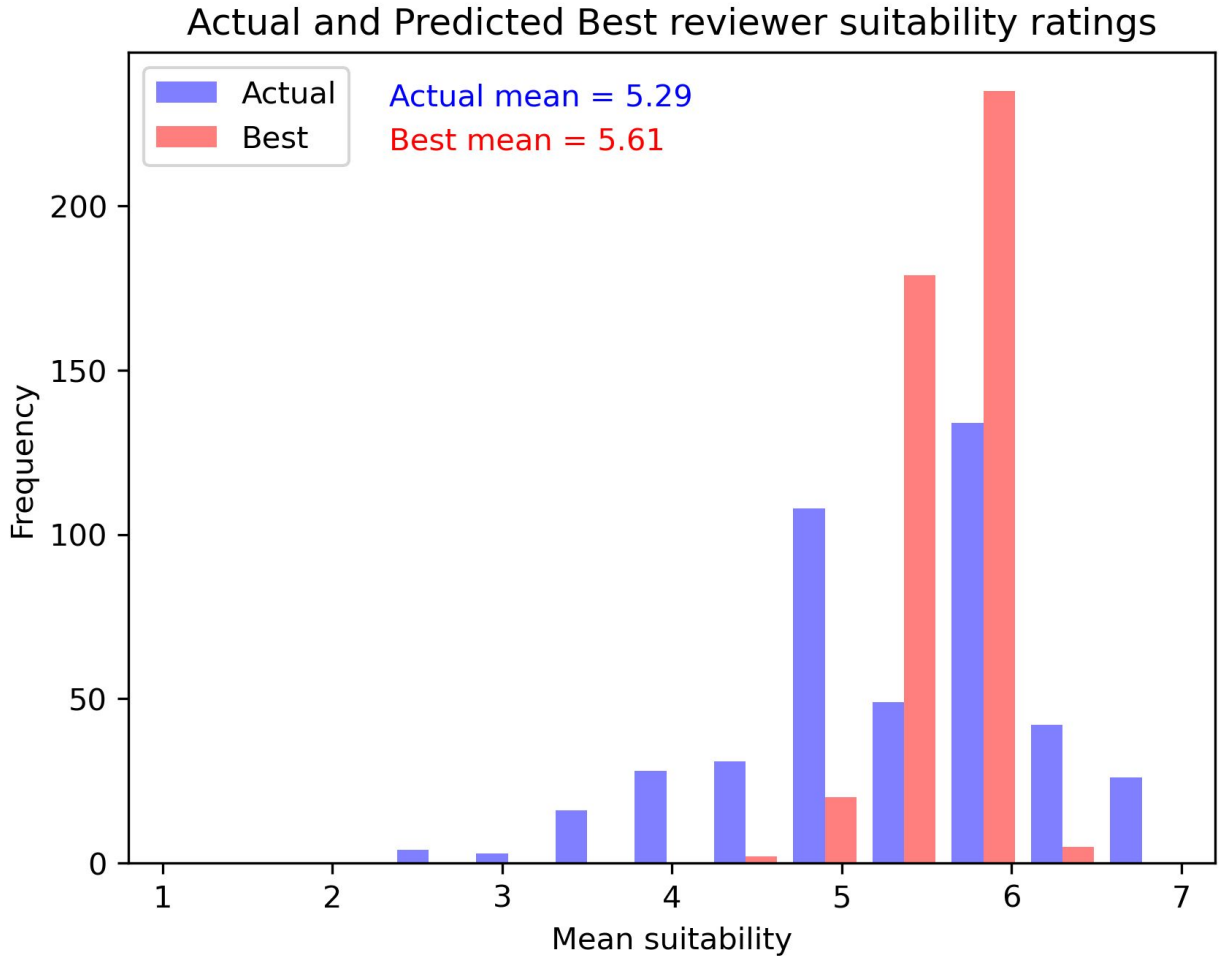
Model: SPECTER2: BERT model pre-trained on scientific texts and augmented by a citation graph



You can predict suitability from matcl



…and from this you can predict gain in suitability from using the optimal match



Research Questions

1. Can language models help match proposals to reviewers?

2. Is it feasible for something like a conference to adopt/adapt this technology?

3. Can it be done securely/privacy respecting?

→ Maybe - evidence for meaningful improvements beyond human matching

→ Definitely yes

→ Definitely yes

Thanks to all participants!