RoRI RESEARCH ON RESEARCH INSTITUTE

AFIRE Sprint on AI in Grantmaking

2025-05-27

Welcome From Tom Stafford

Professor of Cognitive Science
& University Research Practice Lead
University of Sheffield
https://tomstafford.github.io/

Senior Research Fellow,
Research on Research Institute
https://researchonresearch.org/

Welcome From James Wilsdon

Executive Director, RoRI

James was one of the founders of RoRI and has been its director since 2019. He is also Professor of Research Policy at University College London (UCL), based in its Department of Science, Technology, Engineering & Public Policy (STEaPP). Since the late-1990s, as a researcher, writer, adviser and campaigner, James has worked at the heart of science and research policy in the UK and internationally.

You: teams from 22 funders

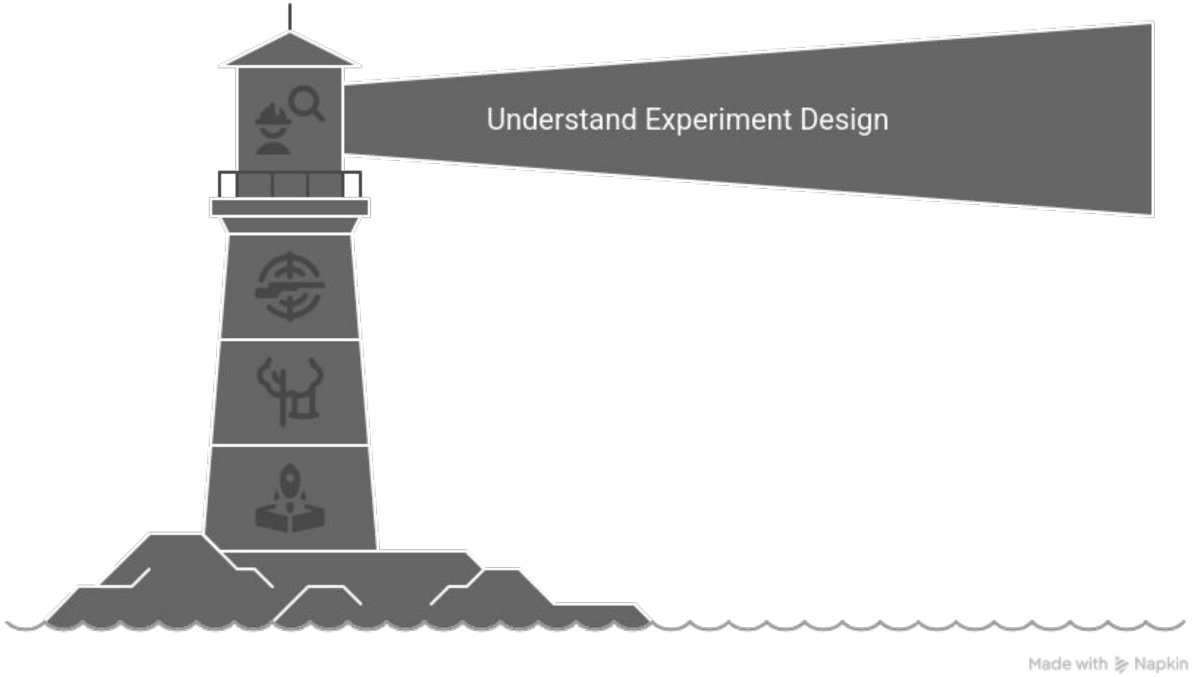
- AQuAS
- CDTI
- CIFAR
- CIHR
- DSIT / Metascience Unit
- FFG
- FNR
- FWF
- HealthResearchBC
- KBF
- La Caixa
- MRC / UKRI
- NSERC
- NWO
- RCN
- Research England
- SNSF
- SSHRC
- Ukrainian Ministry of Education and Science
- Volkswagen Foundation
- Wellcome
- ZonMw

= 22 missions

Introduce yourself in the chat!
+ add your organisation to your zoom name please

Our Mission!

Together we will wrestle with the particular discipline that experiment design forces: how to plan an investigation that is both focussed enough to provide a clear outcome, but also general enough to provide relevant evidence for future action. The ambition is that everyone will finish the sessions some steps closer to the goal of launching new experiments in the space of evaluating the use of AI in research funding.



Roadmap

DATE	FOCUS
May 27	Workshop 1: Why experiment with AI? *Lessons from the GRAIL project, understanding the risks and benefits of experiments*
June 3	- one to one slots : bespoke coaching on developing experiments
June 10	Workshop 2: Examples of experiments with AI *Discussion of case studies, experiment design*
June 17	- one to one slots : bespoke coaching on developing experiments
June 24	Workshop 3: Your experiment with AI *Pitches for new experiments & feedback, advocating for experiments in your organisation*

Some logistics

Email list: you should be on this, let me know if not

Recordings: The talks (only) will be recorded

Discussions are under the Chatham House Rule:

> "anyone who comes to a meeting is free to use information from the discussion, but is not allowed to reveal who made any particular comment."

Coaching sessions: Lead for your group has been contacted / will be soon

Notes file: contains all essential info

Timetable for today

1400 Introductions, Logistics

1410 GRAIL reflections (Youyou Wu)

- **1420** Breakout groups
- **1450** Plenary

1455 (break)

1500 Why experiments? (Tom Stafford)

1555 END

Welcome From Youyou Wu

Associate Professor in Psychology
University College London (UCL)
https://www.drwuyouyou.com/

Research Fellow
Research on Research Institute
https://researchonresearch.org/

GRAIL reflections

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

The Research Council of Norway

wellcome

Swiss National Science Foundation

UKRI Research England

NWO

novo nordisk fonden

Australian Government Australian Research Council

VolkswagenStiftung

DFG Deutsche Forschungsgemeinschaft German Research Foundation

FWF Austrian Science Fund

Swedish Research Council

"la Caixa"

SSHRC CRSH

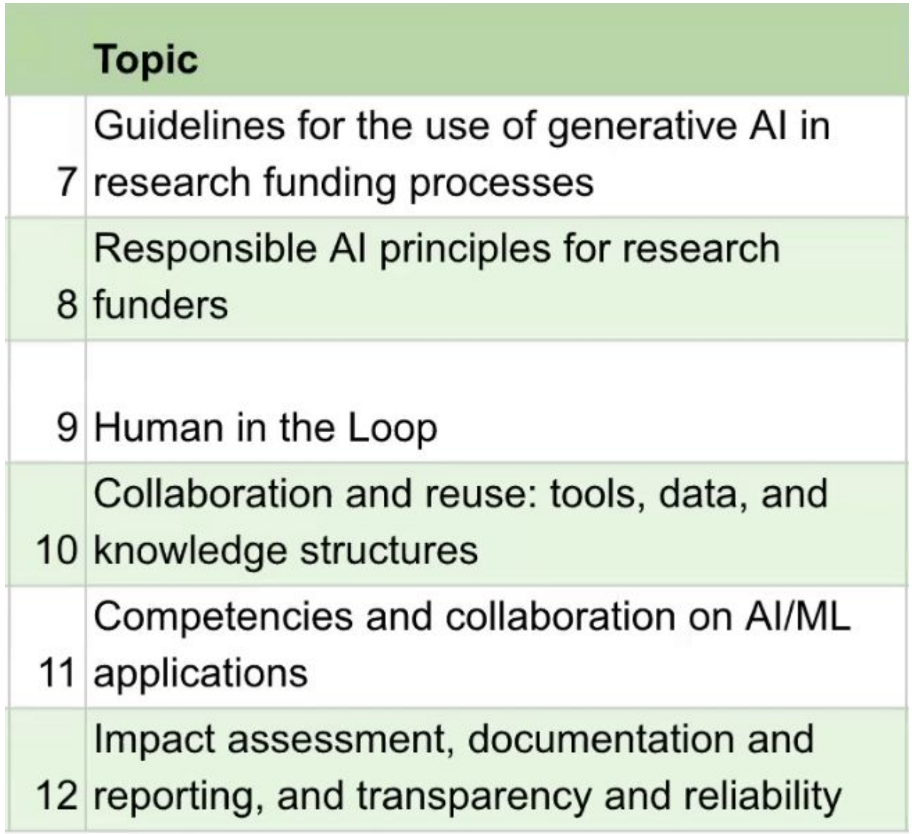
GRAIL activities

13 virtual workshops

Presentations / case studies

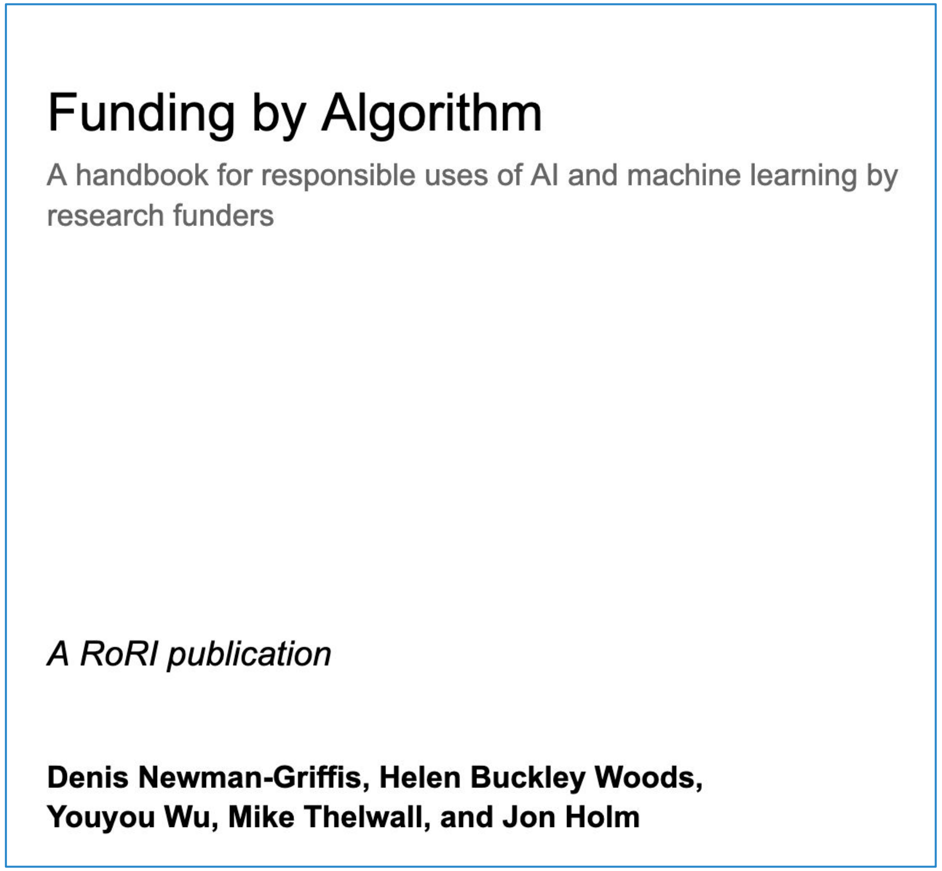
Q&A

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



Funding by Algorithm

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

A RoRI publication

Denis Newman-Griffis, Helen Buckley Woods, Youyou Wu, Mike Thelwall, and Jon Holm

The GRAIL Handbook

[link] *Draft version. Please do not circulate yet*

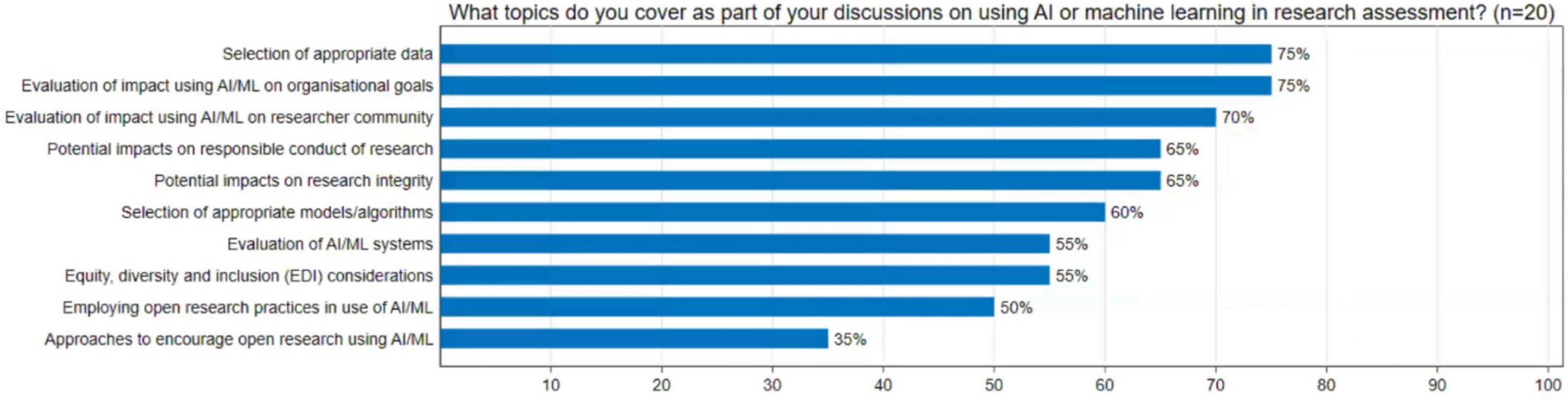
Online launch 20th of June

In-person launch at Metascience conference 2025

GRAIL outputs

A curated library of AI/ML use cases

Survey: experiences with AI/ML



GRAIL insights

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: AI/ML development

AI/ML development cycles

Prototyping
first version, see if it can be done

Pilot testing
apply to a small set of cases

Beta testing and refinement
gradually roll out to users

Organisational deployment
fully deployed, user training and monitoring

GRAIL insights: AI/ML development

AI/ML development cycles

Prototyping
first version, see if it can be done

Pilot testing
apply to a small set of cases

Beta testing and refinement
gradually roll out to users

Organisational deployment
fully deployed, user training and monitoring

AI/ML development procedures

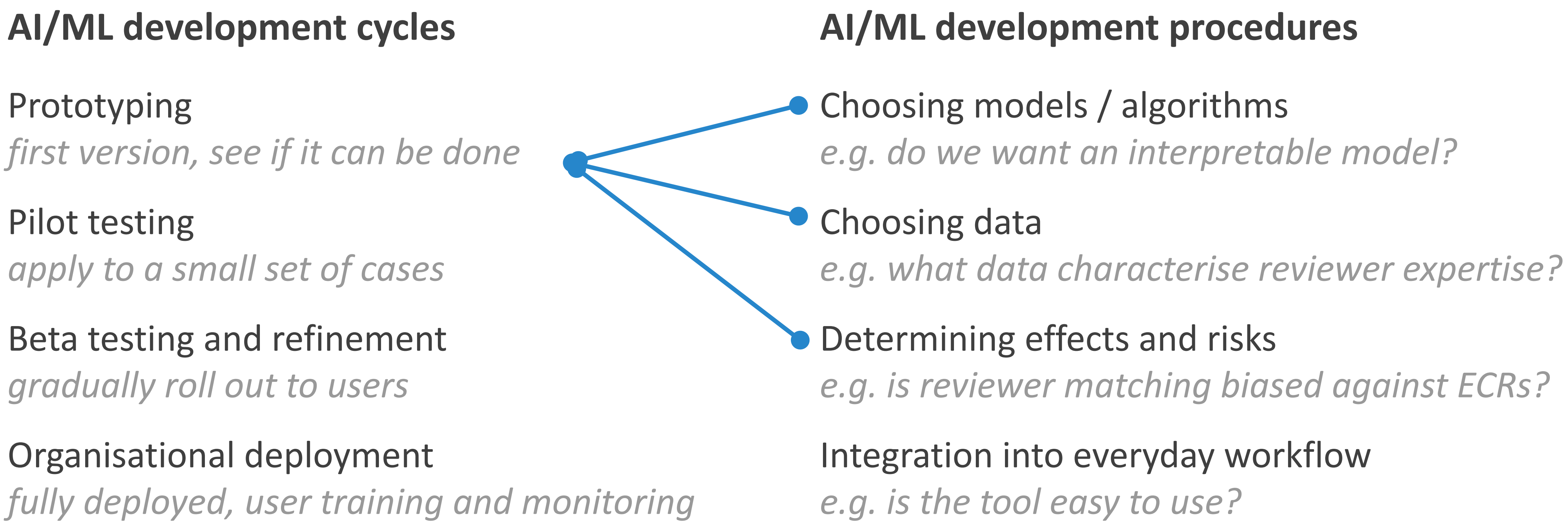
Choosing models / algorithms
e.g. do we want an interpretable model?

Choosing data
e.g. what data characterise reviewer expertise?

Determining effects and risks
e.g. is reviewer matching biased against ECRs?

Integration into everyday workflow
e.g. is the tool easy to use?

GRAIL insights: AI/ML development



AI/ML development cycles

Prototyping
first version, see if it can be done

Pilot testing
apply to a small set of cases

Beta testing and refinement
gradually roll out to users

Organisational deployment
fully deployed, user training and monitoring

AI/ML development procedures

Choosing models / algorithms
e.g. do we want an interpretable model?

Choosing data
e.g. what data characterise reviewer expertise?

Determining effects and risks
e.g. is reviewer matching biased against ECRs?

Integration into everyday workflow
e.g. is the tool easy to use?

GRAIL insights: AI/ML development

AI/ML development cycles

Prototyping
first version, see if it can be done

Pilot testing
apply to a small set of cases

Beta testing and refinement
gradually roll out to users

Organisational deployment
fully deployed, user training and monitoring

AI/ML development procedures

Choosing models / algorithms
e.g. do we want an interpretable model?

Choosing data
e.g. what data characterise reviewer expertise?

Determining effects and risks
e.g. is reviewer matching biased against ECRs?

Integration into everyday workflow
e.g. is the tool easy to use?

Evaluation and experimentation can happen at all the stages, across all the procedures

GRAIL insights: evaluation and experimentation

Why do formal experiments?

Validate the AI/ML tools

Align with organisational goals

Gain trust and support from applicants / institutions / the public

Set standards for the wider research community

GRAIL insights: evaluation and experimentation

Why do formal experiments?

Validate the AI/ML tools

Align with organisational goals

Gain trust and support from applicants / institutions / the public

Set standards for the wider research community

What to experiment on?

Does the AI/ML tool work?

Not just "does it work", but "How, for whom, and with what consequences"?

AI implementation is never "done": Continuous monitoring, refinement, and feedback loops

GRAIL insights: evaluation and experimentation

Why do formal experiments?

Validate the AI/ML tools

Align with organisational goals

Gain trust and support from applicants / institutions / the public

Set standards for the wider research community

What to experiment on?

Does the AI/ML tool work?

Not just “does it work”, but “How, for whom, and with what consequences”?

AI implementation is never “done”: Continuous monitoring, refinement, and feedback loops

Just like how you would evaluate any other non-AI processes

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

Breakout group discussion

Topic: Your experiences working on AI/ML applications
Duration: 30 minutes
Structure:

Going around each funder:

- What AI/ML application have you worked on?
- At what stage are you with it? (prototyping, pilot testing, beta testing and refinement, full deployment)
- Where are you with experimentation? Tag yourself with the following:
  Experimenter (have done experiments to evaluate the AI/ML tool)
  Designer (have designed experiments, with implementation soon to follow)
  Thinker (thinking about experimenting, but are yet to take further steps)

For each funder, the aim is to have **a short sentence to share at the end explaining what experiments you have done/designed/thought about**.

Plenary:

Please share in the chat:

a short sentence explaining what experiments you have done/designed/thought about

Break

Start again on the hour (1600 CEST)

Feedback from pre-workshop survey

1. Which of these is your team most interested in working on?

2. In a few short words, what would be the ideal outcomes for you of these sessions?

3. Is there anything else you think we should know?

most interested in working on?

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

ideal outcomes for you of these sessions?

Ideas and plans for **experiments**

Collaborative leaning: pulse check, see tools and approaches other funding agencies are using; identify most promising uses of AI, pitfalls, feedback or concerns; share best practices; hear concerns of their staff and stake-holders

Tools: identifying suitable reviewers; stardarized added-value reports ; workflow efficiencies

Techincal understanding: learning materials for building local LLMs, time and resources required for implementaiton and deployment.

anything else you think we should know?

Individual queries: let me know if you haven’t heard from me

Note: different priorities across organisation

Other uses: making science more impactful; assess peer review process (e.g. detect biases)

Other interests: risks of AI that funders need to mitigate; privacy and ethical concerns; the how of AI (e.g. using models in safe and close infrastructure).

Why experiment with AI?

Better evidence: Why innovation & observation alone are not enough

Evidence > Experience

Ignaz Philipp Semmelweis (1818-1865)
Universal Images Group via Getty Images

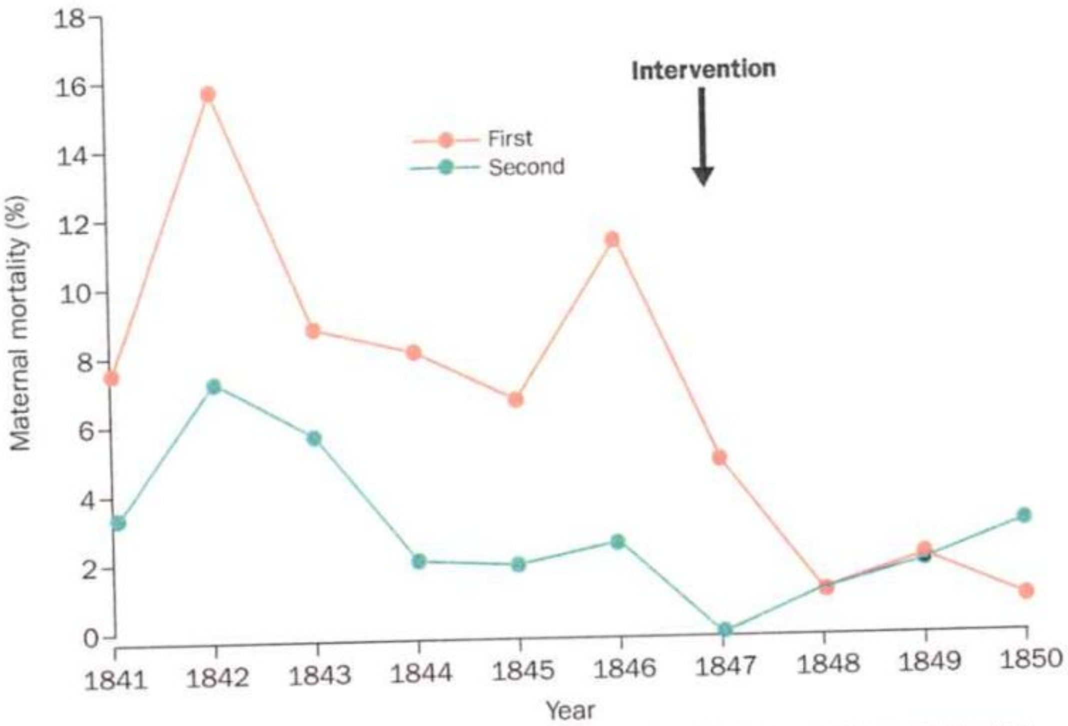


Figure 2. Maternal mortality rates in the First and Second Clinic at the Lying-In Women's Hospital, Vienna, before and after hand hygiene in chlorinated lime had been introduced in May, 1847. Rates have been calculated according to numbers given in reference 22.

Pittet, D., & Boyce, J. M. (2001). Hand hygiene and patient care: pursuing the Semmelweis legacy. *The Lancet Infectious Diseases*, *1*, 9-20.

Experiments > Observation

out of 52 claims about nutrition based on observational studies, none replicated in randomised trials (and 5 trials showed effects in the opposite direction)

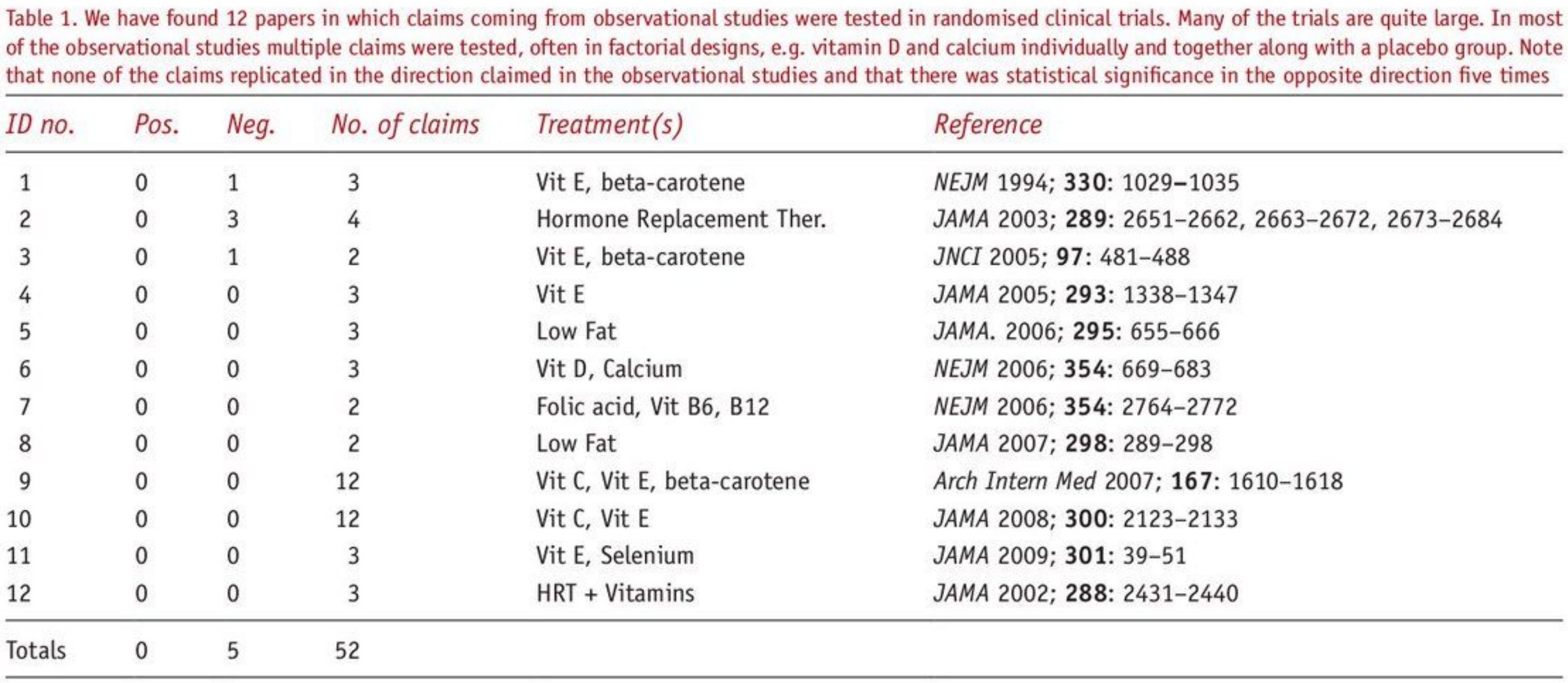


Table 1. We have found 12 papers in which claims coming from observational studies were tested in randomised clinical trials. Many of the trials are quite large. In most of the observational studies multiple claims were tested, often in factorial designs, e.g. vitamin D and calcium individually and together along with a placebo group. Note that none of the claims replicated in the direction claimed in the observational studies and that there was statistical significance in the opposite direction five times

ID no.	*Pos.*	*Neg.*	*No. of claims*	*Treatment(s)*	*Reference*
1	0	1	3	Vit E, beta-carotene	*NEJM* 1994; **330**: 1029–1035
2	0	3	4	Hormone Replacement Ther.	*JAMA* 2003; **289**: 2651–2662, 2663–2672, 2673–2684
3	0	1	2	Vit E, beta-carotene	*JNCI* 2005; **97**: 481–488
4	0	0	3	Vit E	*JAMA* 2005; **293**: 1338–1347
5	0	0	3	Low Fat	*JAMA*. 2006; **295**: 655–666
6	0	0	3	Vit D, Calcium	*NEJM* 2006; **354**: 669–683
7	0	0	2	Folic acid, Vit B6, B12	*NEJM* 2006; **354**: 2764–2772
8	0	0	2	Low Fat	*JAMA* 2007; **298**: 289–298
9	0	0	12	Vit C, Vit E, beta-carotene	*Arch Intern Med* 2007; **167**: 1610–1618
10	0	0	12	Vit C, Vit E	*JAMA* 2008; **300**: 2123–2133
11	0	0	3	Vit E, Selenium	*JAMA* 2009; **301**: 39–51
12	0	0	3	HRT + Vitamins	*JAMA* 2002; **288**: 2431–2440
Totals	0	5	52		

Deming, data and observational studies (2011)
https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1740-9713.2011.00506.x

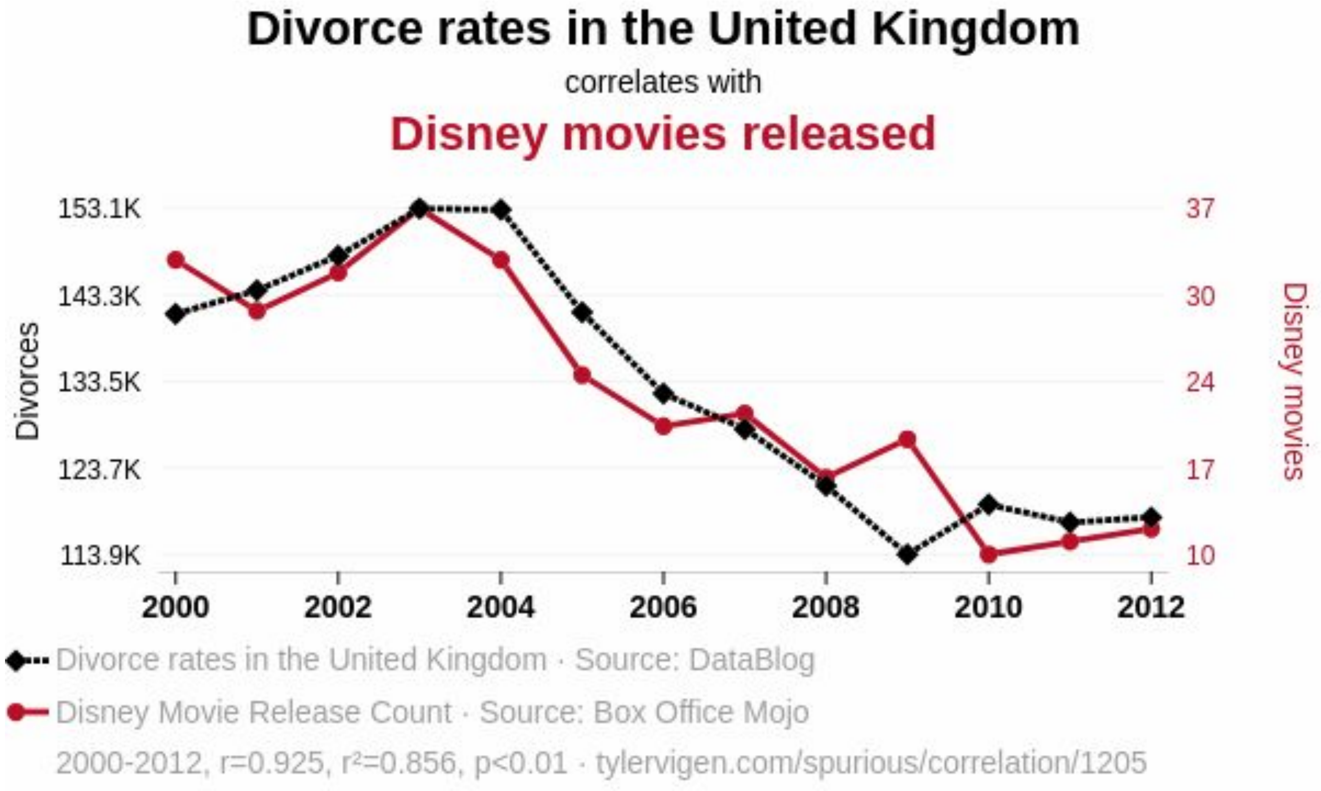
Limits to observation

Spurious correlations

Omitted variables

(confounding)

Out of sample



2000-2012, r=0.925, r^2=0.856, p<0.01 · tylervigen.com/spurious/correlation/1205

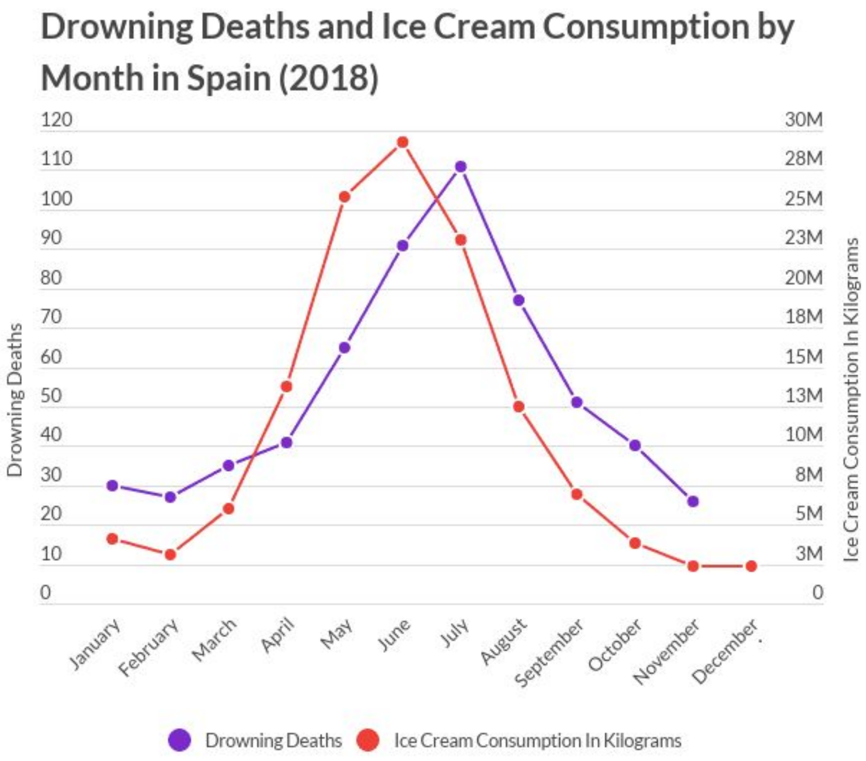
Limits to observation

Spurious correlations

Omitted variables

(confounding)

Out of sample



Statista (2020)

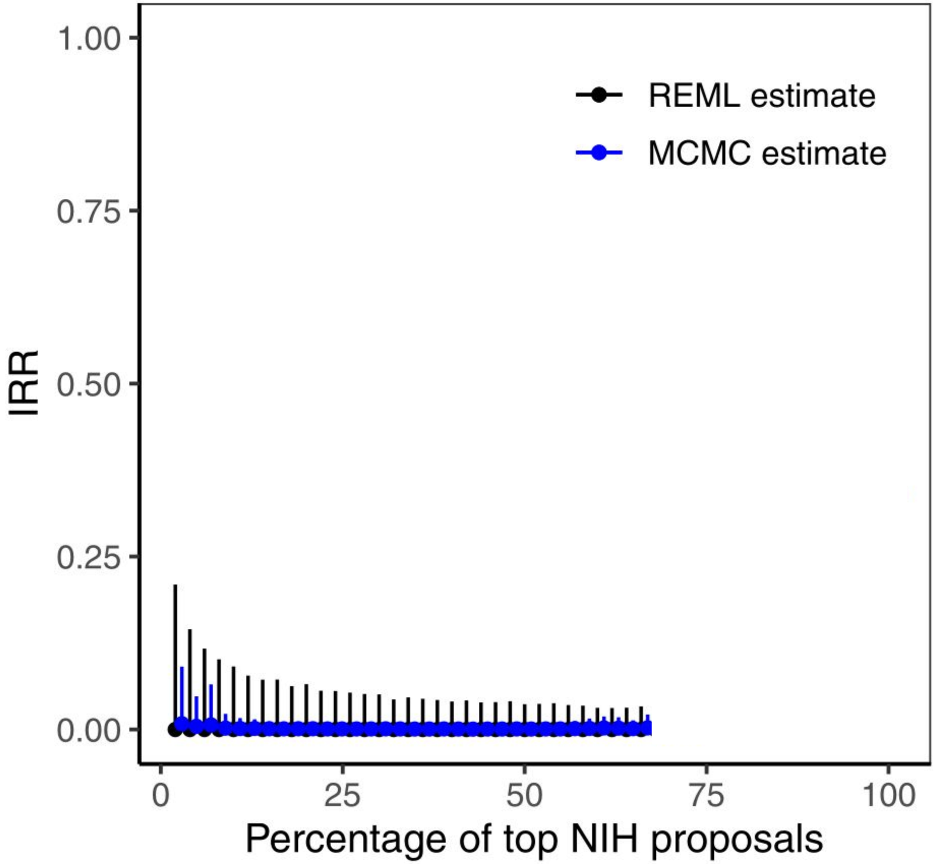
Limits to observation

Spurious correlations

Omitted variables

(confounding)

Out of sample



Erosheva EA, Martinková P, Lee CJ. When zero may not be zero: A cautionary note on the use of inter- rater reliability in evaluating grant peer review. J R Stat Soc Series A. 2021;00:1–16. https://doi.org/10.1111/rssa.12681

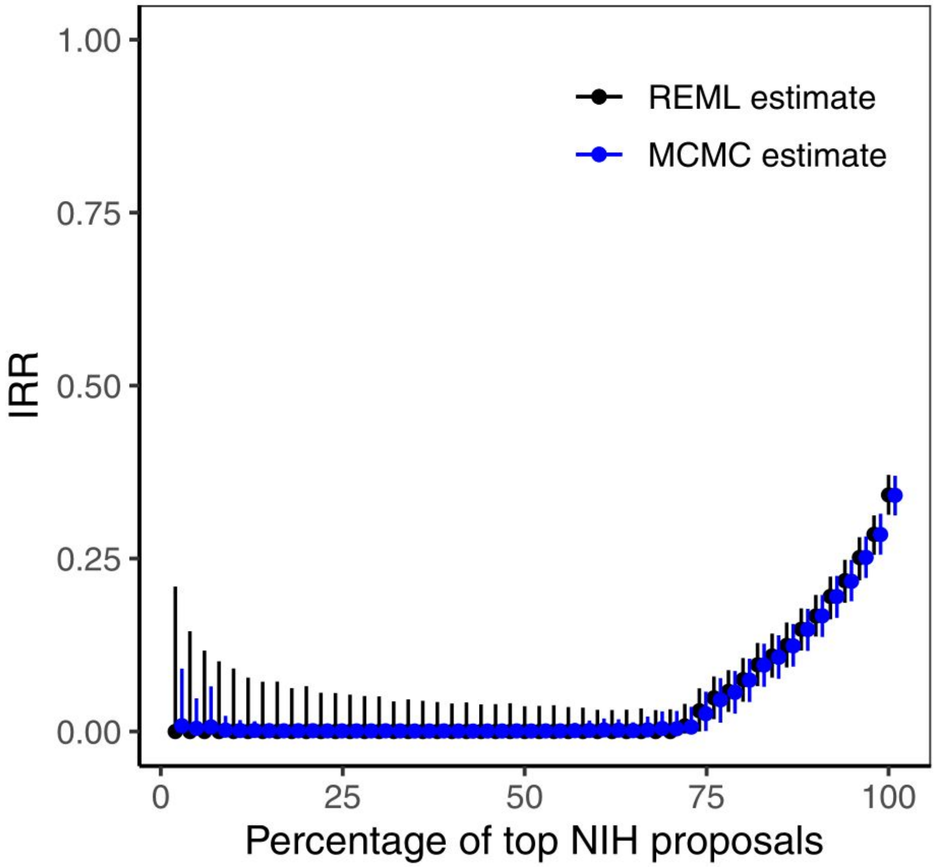
Limits to observation

Spurious correlations

Omitted variables

(confounding)

Out of sample



Erosheva EA, Martinková P, Lee CJ. When zero may not be zero: A cautionary note on the use of inter- rater reliability in evaluating grant peer review. J R Stat Soc Series A. 2021;00:1–16. https://doi.org/10.1111/rssa.12681

What “Experiment” means

A change or difference ->

A commitment to evaluate

A plan to evaluate

A commitment to share the evaluation

Analysis plan is essential

Prespecification - in as much detail as possible:

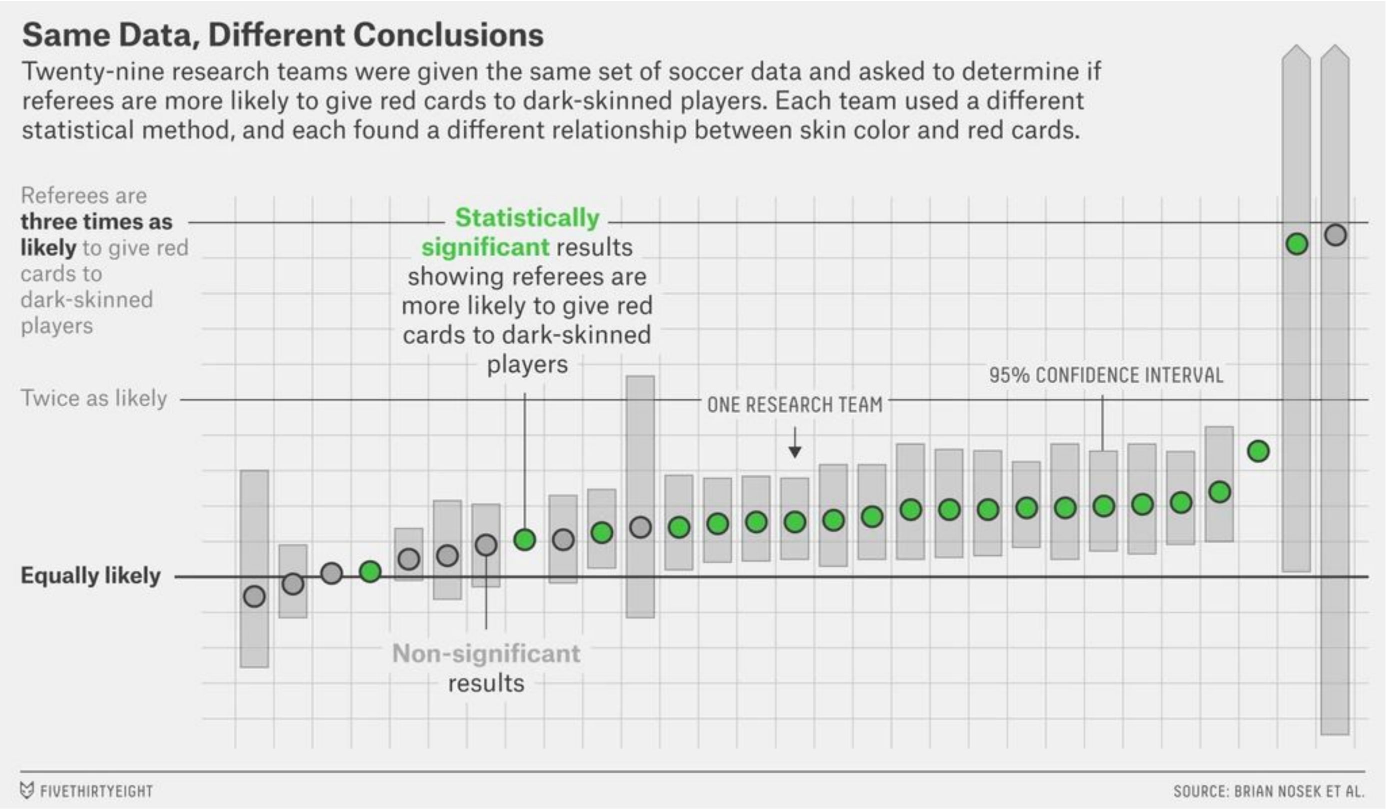
- The research question(s)

- The primary outcome measure(s)

- What you’re going to do (which interventions, who with / which data)

- Test(s) which will determine success

Researcher degrees of freedom



Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

Statistically significant results showing referees are more likely to give red cards to dark-skinned players

Non-significant results

ONE RESEARCH TEAM

95% CONFIDENCE INTERVAL

FIVETHIRTYEIGHT

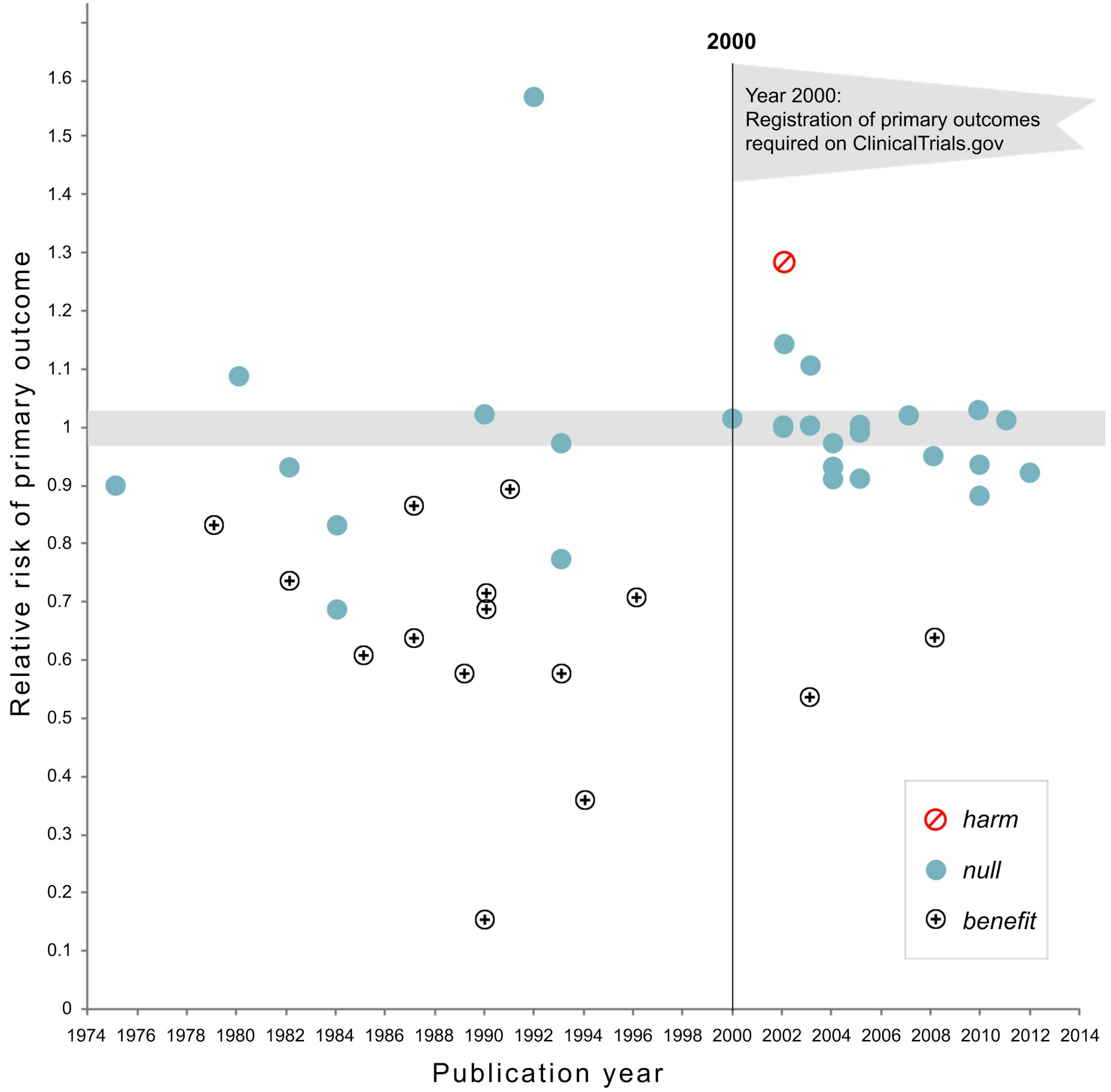
SOURCE: BRIAN NOSEK ET AL.

Referees are **three times as likely** to give red cards to dark-skinned players
Twice as likely
Equally likely

"These findings suggest that significant variation in the results of analyses of complex data may be difficult to avoid, even by experts with honest intentions."

Silberzahn, R. et al (2018). [Many analysts, one dataset: Making transparent how variations in analytical choices affect results](). *Advances in Methods and Practices in Psychological Science, 1*(3), 337-356

Analytic flexibility can mislead



Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PloS one*, *10*(8), e0132382. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132382

Sharing is essential

Contribute to knowledge commons

Support future evidence synthesis

Advertise your work

Transparency of signal of credibility

Threat of scrutiny forces rigour

Learning organisations - requires bravery!

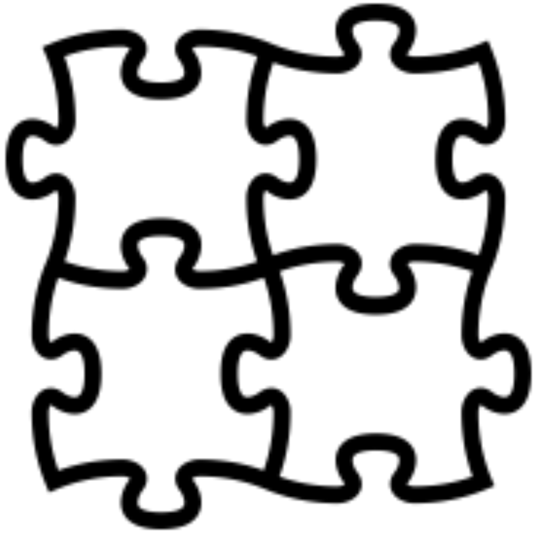


Image: Puzzle Pieces, CC-BY by M Ryan, US

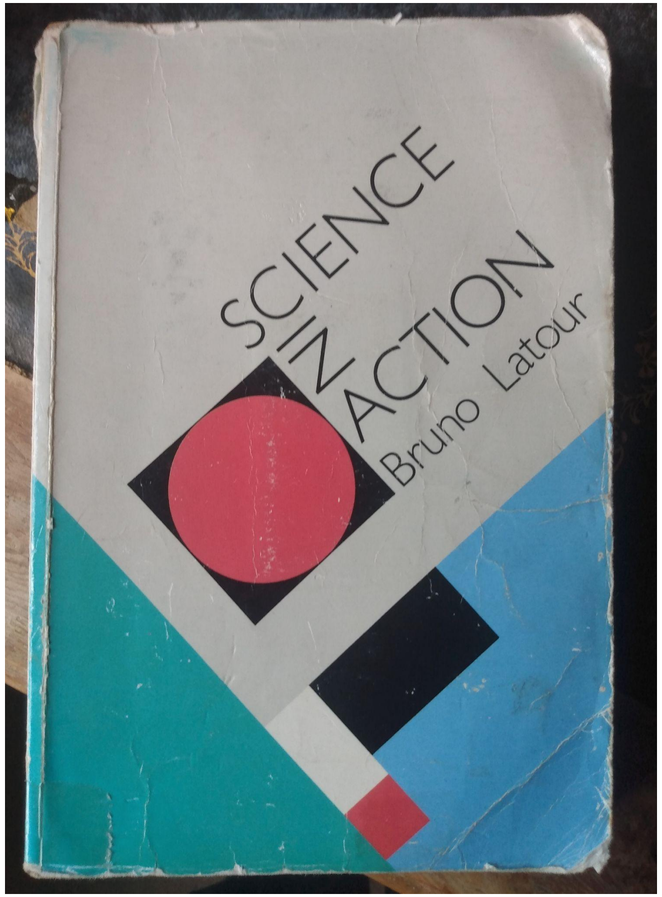
When to experiment

Show values

Build learning capacity: how to explore?

Resolve uncertainty : how to exploit?

Propagandise change



Assays and microscopes

Yes/No

but is it the right question

Close view

but what are you looking for?

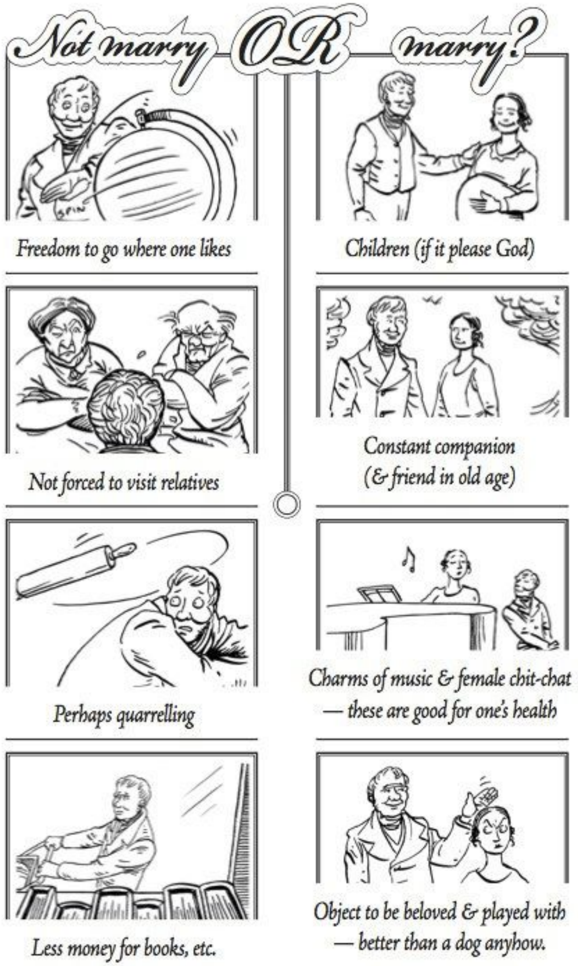
Image: CC Wikimedia

Just designing experiments is valuable

Marry:
Children—(if it Please God) — Constant companion, (& friend in old age) who will feel interested in one,— object to be beloved & played with.— better than a dog anyhow.— Home, & someone to take care of house— Charms of music & female chit-chat.— These things good for one's health.

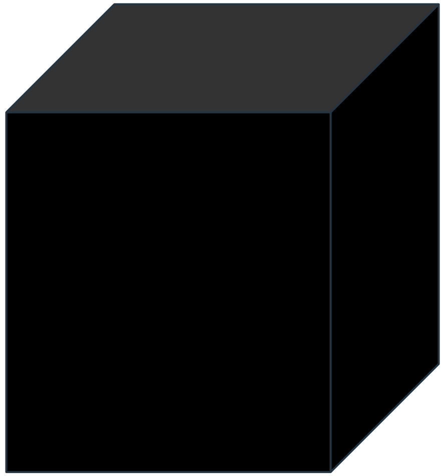
Not Marry:
Freedom to go where one liked— choice of Society & little of it. — Conversation of clever men at clubs— Not forced to visit relatives, & to bend in every trifle.— to have the expense & anxiety of children— perhaps quarelling— Loss of time. — cannot read in the Evenings— fatness & idleness— Anxiety & responsibility— less money for books &c— if many children forced to gain one's bread.

Image: Darwin: A Graphic Biography Paperback (2013) Eugene Byrne & Simon Gurr


Not marry OR marry?
Freedom to go where one likes
Children (if it please God)
Not forced to visit relatives
Constant companion (& friend in old age)
Perhaps quarrelling
Charms of music & female chit-chat — these are good for one's health
Less money for books, etc.
Object to be beloved & played with — better than a dog anyhow.

Why AI specifically requires experiments

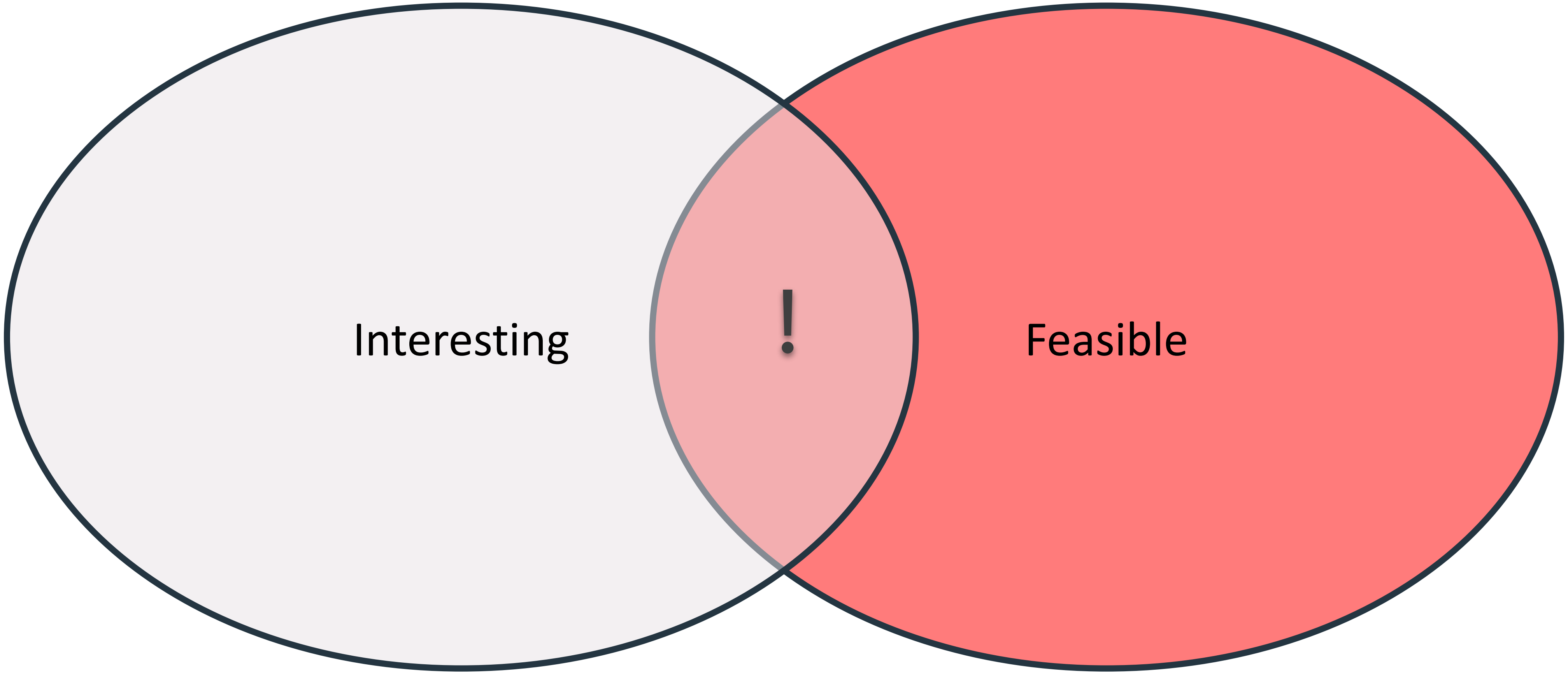
AI is novel & changing - our intuitions are poorly calibrated



Specific areas of uncertainty

- Fluency
- Stochasticity
- Hallucinations
- Inscrutability
- Bias

“The Art of the Possible”



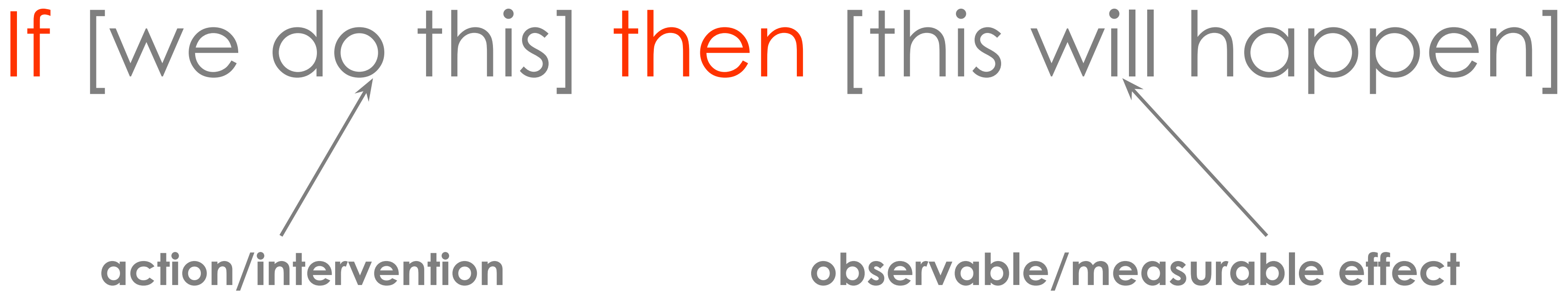
Good outcome measures

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

~~IDEA → SOLUTION~~

IDEA → HYPOTHESIS

If … then…



If [we do this] then [this will happen]

action/intervention

observable/measurable effect

Breakout groups until h50 mins

If … then…

If [we do this] then [this will happen]

Reformulate AI use-case from first break out as an “if-then” statement

“If X then Y”

“If we could use AI to more quickly identify reviewers then we could make awards more quickly after application”

Plenary: 1640 CEST

i) THEN : What targets did we have for our use of AI?

ii) IF : What challenges did we identify for interventions?

Next time

Homework: Which ONE use of AI in your organisation is easiest to test if it is working (and HOW). Formulate as an IF-THEN statement.

Session 1 feedback survey

Pre-session activity for session 2: coming soon!

One-to-one coaching slot: 3rd of June (or anytime that week)

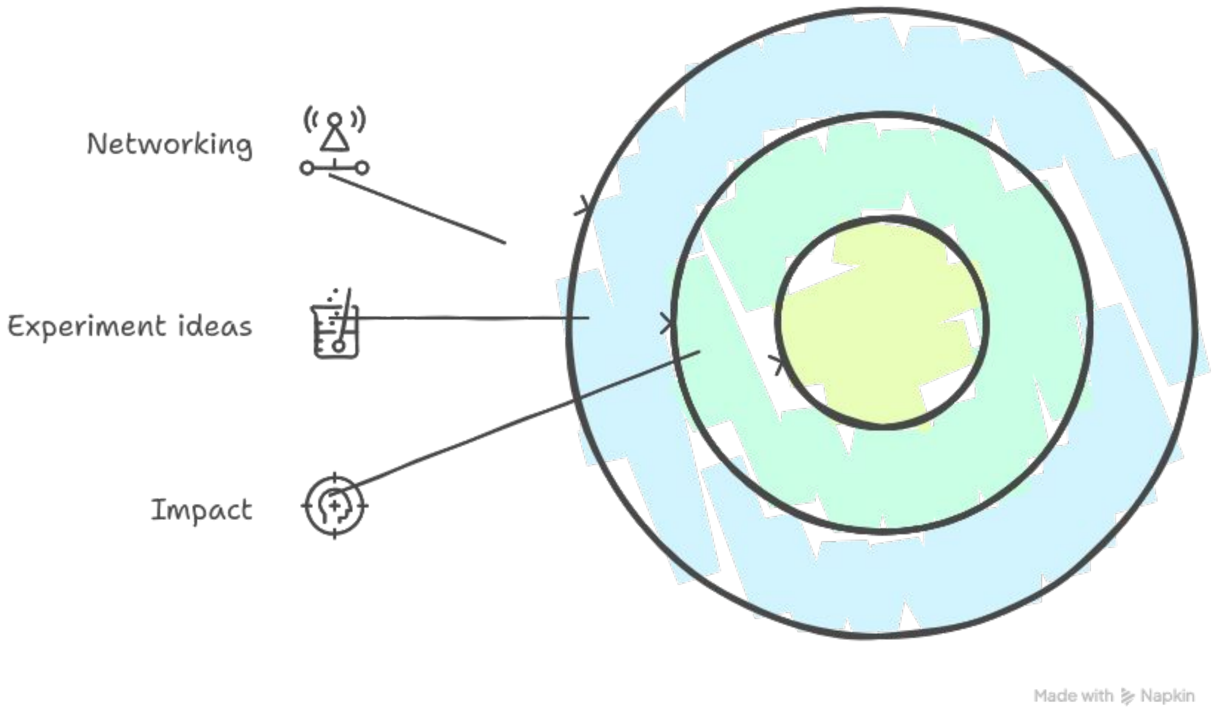
Next workshop: 10th of June: case studies in experiments with AI + fundamentals of experiment design

METASCIENCE 2025 CONFERENCE

Save the Date

June 30 – July 2, 2025
University College London

Metascience Lab @ MS2025



- in partnership with Open Philanthropy and RoRI’s AFiRE programme

- three linked sessions will facilitate matchmaking and networking for experimentation

- all areas of metascience, with a focus on interventions to support higher quality, lower cost and more impactful research.

- Each session will showcase metascience principles, methods or examples of experimentation, as well as providing a platform for co-developing new project ideas by participants. Researchers, funders, universities, publishers and other actors in the research ecosystem are invited to propose experiments and matchmake with potential collaborators.

- The Abundance and Growth Fund at Open Philanthropy is happy to consider proposals that emerge from this process

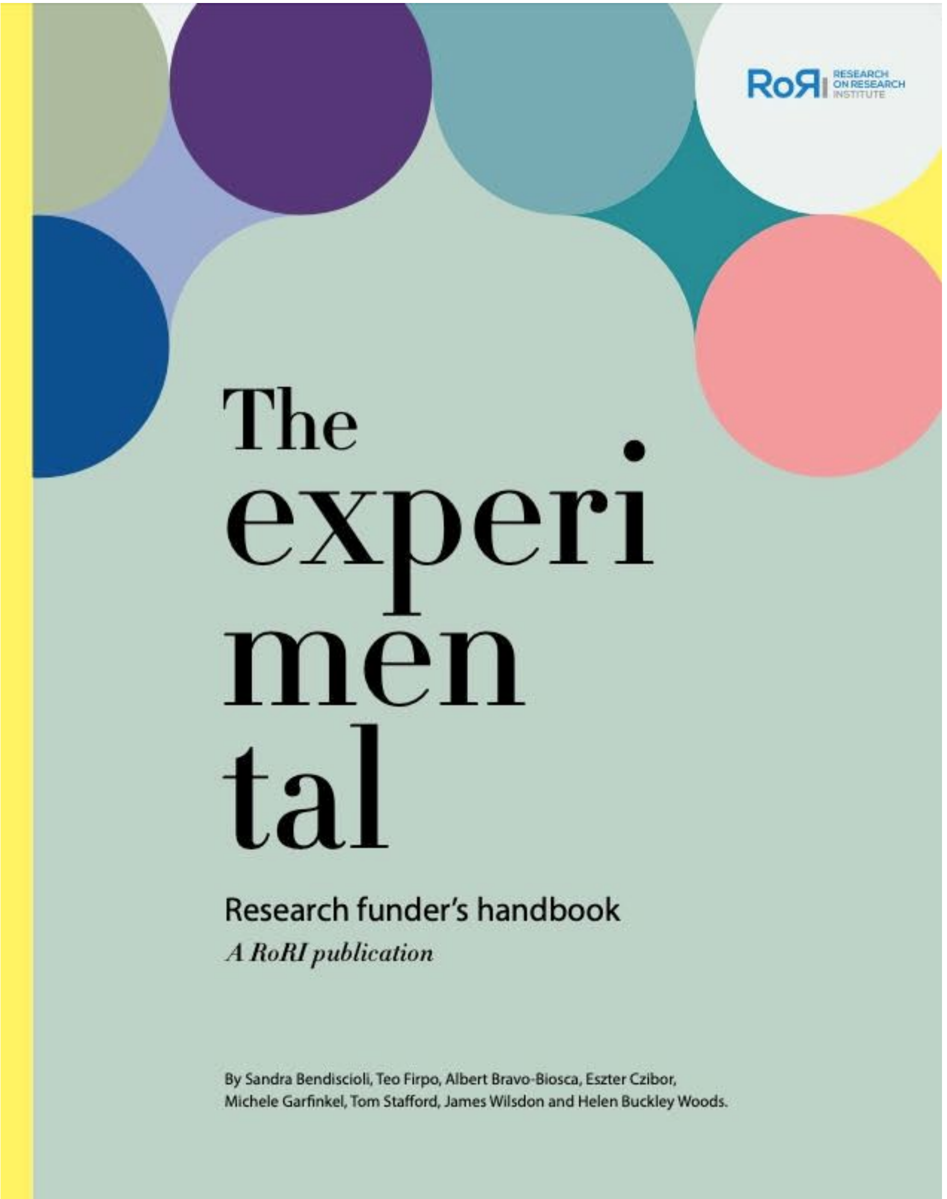
- Topics you’d like considered? Please get in touch

Art of Funding @ MS2025

Come join a small group of funders to discuss the “Art of Funding”. Topics may include advancing new ideas within your organization, overcoming bottlenecks, efficiencies, and logistics of making and monitoring awards.

Bring your questions and ideas to share. **Drinks will be served**.

Please RSVP https://forms.gle/aHhpVWWc2gs7Fpux8. Discussions will continue afterwards at a restaurant of your choice.



RoRI RESEARCH ON RESEARCH INSTITUTE

The experi men tal

Research funder's handbook

A RoRI publication

By Sandra Bendiscioli, Teo Firpo, Albert Bravo-Biosca, Eszter Czibor, Michele Garfinkel, Tom Stafford, James Wilsdon and Helen Buckley Woods.

Bendiscioli, Sandra; Firpo, Teo; Bravo-Biosca, Albert; Czibor, Eszter; Garfinkel, Michele; Stafford, Tom; et al. (2022):

The experimental research funder's handbook (Revised edition, June 2022, ISBN 978-1-7397102-0-0).

Research on Research Institute. Report. https://doi.org/10.6084/m9.figshare.19459328.v2

Further resources on experimentation

Bendiscioli, Sandra; Firpo, Teo; Bravo-Biosca, Albert; Czibor, Eszter; Garfinkel, Michele; Stafford, Tom; et al. (2022): The Experimental Research Funder's Handbook (Revised edition, June 2022, ISBN 978-1-7397102-0-0). Research on Research Institute.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. Nature human behaviour, 1(1), 1-9. https://doi.org/10.1038/s41562-016-0021

Why observation is not enough, even if you have sophisticated analysis:

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. Marketing Science, 38(2), 193-225.

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. PLOS ONE, 11(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Join the conversation

researchonresearch.org

@RoRInstitute

RoRI RESEARCH ON RESEARCH INSTITUTE