# We Need a Rigorous Metascience of Artificial Intelligence

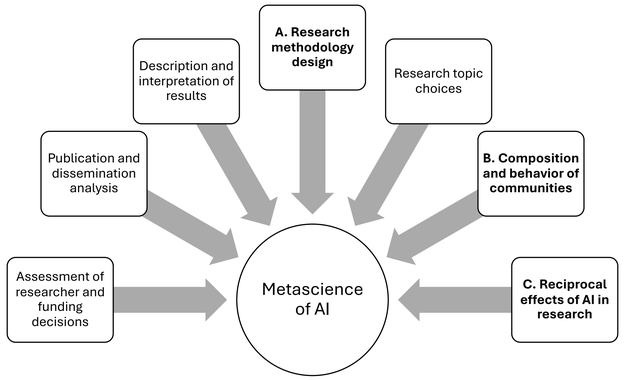
*Bálint Gyevnár – Carnegie Mellon University*

Carnegie Mellon University

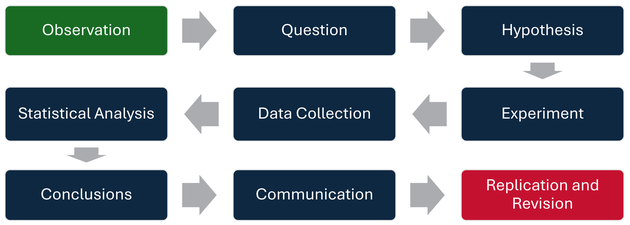
## What is the metascience of AI?

*The rigorous scientific study of:*

1. AI research process,
2. Effects of AI on research itself.




*Using the scientific process:*




## Why is the metascience of AI crucial?

*There are at least 10 reasons from AI research:*

1. p-hacking is a feature not a bug,
2. Benchmarks don't correlate with capabilities,
3. No way to distinguish hype from substance,
4. Peer review is no better than random,
5. Overwhelming AI-enabled slop,
6. Emergence of epistemic monocultures,
7. Biased and unsafe AI steers science
8. Scientist AIs are demonstrably flawed,
9. Rampant irreproducible research,
10. Constantly reinventing the wheel.

Using biased and unsafe AI for science may be steering human research efforts towards epistemic and procedural disintegration.

A rigorous metascience of AI can provide the tools and experiments necessary to understand and mitigate the harmful effects.

## How can metascience of AI help?

*(biased selection of examples)*

*Normative studies with empirical results:*

- Expanding epistemic scope of AI safety,
- Bridging shared research problems among AI safety and AI ethics,
- Pitfalls of trusting scientist AI to do automated research

*Empirical methods for testing AI effect:*

- Differences between human-written mathematical proofs / scientific code / etc. from AI-written ones?
- Effects of AI-use on scientific integrity.

*Rigorous tools to mitigate AI harms:*

- Methodology to elicit bias in controlled experiments,
- Automatic verification of AI-generated research traces,
- Methods to document scientist AI data.

## Case Studies

### A. How do we protect scientific integrity in the age of scientist AIs?

Paper: The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems

*Do Scientist AI systems follow basic research practice?*
Spoiler: No.

Through controlled experiments with a novel task showed:

- **Inappropriate benchmark selection:** prefer benchmarks with high SOTA or choose in order of the candidate list,
- **Data leakage:** often generates and reports results on contrived synthetic data,
- **Metric misuse:** Arbitrary choices sensitive to metric ordering, but no deliberate abuse,
- **Post-hoc selection bias:** systems peak at evaluations on the test set which akin to training on the test set.

### B. How do we bridge the research problems of polarized fields of AI?

Paper: AI Safety for Everyone

*What risks are most discussed in AI safety literature?*
Spoiler: It is much more than X-risks.

Via literature review of primarily peer reviewed publications, we identify three risks of equating AI safety with X-risks:

- **Exclude researchers** who are committed to AI safety but approach the field from different angles,
- **Lead the public to mistakenly view AI safety** as focused solely on existential scenarios rather than addressing a wide spectrum of safety challenges,
- **Risks creating resistance to safety** measures among those who disagree with predictions of existential AI risks

### C. What are the differences between human and AI research?

Post: Mathematical Understanding and Artificial Intelligence (WIP)

*How does AI affect the way mathematicians prove?*
Spoiler: AI is much more reckless changing proofs.

Through a comparison of Lean4 GitHub repository commits to AI-suggested changes we find:

- **Humans have stable ontologies:** once a function refers to many other functions, or is used by many other functions, people don't change it
- **LLMs disrupt existing proofs:** readily introduce high-level methods or rewrite very low-level proofs,