

The Truth, the Whole Truth, and Nothing but the Truth: A Pragmatic Guide to Assessing Empirical Evaluations

Stephen M. Blackburn et al.
Published: 13 October 2016

Presented by Brian Do - 2025/09/16

Problem being solved

Motivation:

- Many research papers make unsound or poorly supported claims
- These lead to:
 - Misleading conclusions
 - Waste of community effort
 - Difficulty building on prior work

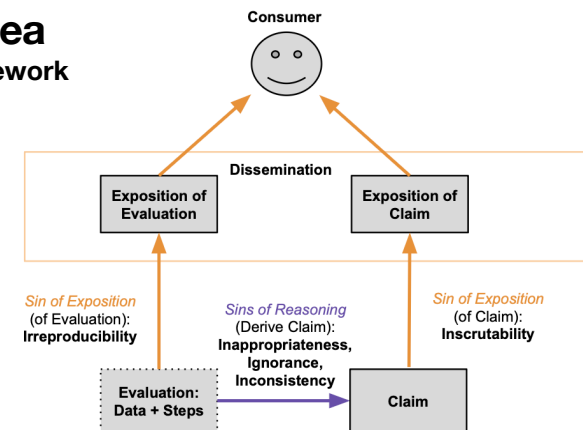
Goal:

- Provide a framework to classify and avoid these mistakes
- Encourage the community to think about sound claims
- Encourage a culture shift toward strong evaluation or high novelty papers

Connection to the book chapter

- Software analysis relies on data-driven claims to inform decisions
- This paper's framework helps ensure:
 - Evidence really supports claims
 - Results are actionable for practitioners

New idea The Framework



Framework explaining why claims become unsound and poorly described

New idea

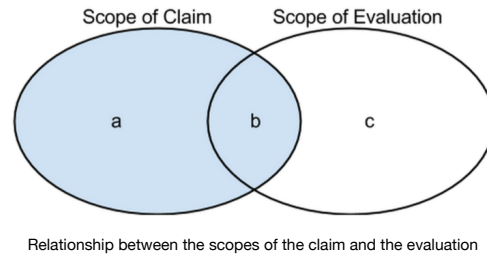
The Framework

Sins of reasoning:

- Ignorance: (a) empty
- Inappropriateness: (c) empty
- Inconsistency: (b) empty

Sins of exposition:

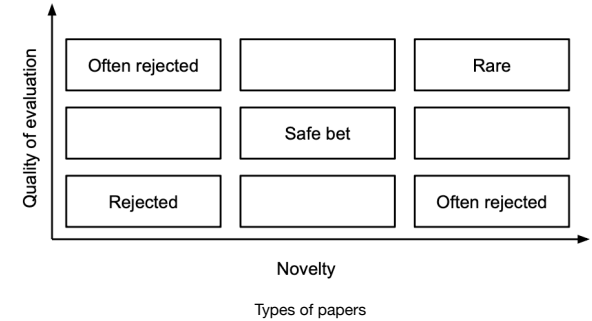
- Inscrutability: inadequate claim
- Irreproducibility: inadequate evaluation



New Idea

Call for culture change

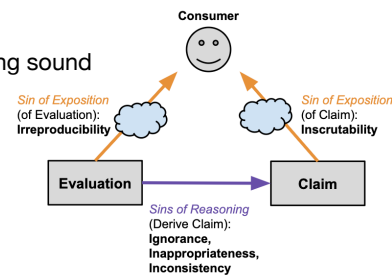
Low-novelty, high-evaluation and high-novelty, low-evaluation papers are important and worth publishing.



Critique

Positive points

- Effective figures illustrate the framework visually
- Provides actionable guidance
- Encourages a cultural shift towards valuing sound claims and evaluation



Critique

Negative points

- Domain-specific examples
- Lacks guidance for prioritizing or weighting sins

Future work

Conducting a study of the distribution of sins across published papers:

- Classify papers by venue or field to see which sins are most common
- Investigate field-specific variations
- Goal: raise awareness & improve publishing/review practices

Reflections

Rating: 4/5

- Valuable for understanding claim–evidence relationships
- Influences how I approach my own research and evaluations

Discussion Points

- Do these “sins” map cleanly to other fields (e.g., HCI, ML, SE)?
- Are some sins more serious or damaging than others?
- Could automated tools be developed to detect ambiguous claims or missing evidence?

**Thank you for listening and
participating!**