CS846 Week 2 Reviews: Paper 2

Brian Do

Paper: *The Truth, the Whole Truth, and Nothing but the Truth: A Pragmatic Guide to Assessing Empirical Evaluation*, Blackburn et al., ACM Trans. on Programming Languages and Systems (TOPLAS), 38(4), Oct. 2016.

Problem Being Solved

The authors aim to address the prevalence of unsound or unclear claims within research papers, and the lack of a shared framework for discussing these problems. While prior work has emphasized the importance of sound empirical evaluation, there has been little guidance on how to reason about the relationship between evidence, claims, and communication. The authors propose to fill this gap by introducing a conceptual framework that names and categorizes the common ways claims can go wrong. This complements the book chapter on software analytics, which emphasizes the need for rigorous, interpretable, and actionable insights when drawing conclusions from data.

New Idea

The paper introduces a principled framework that explains how unsound claims arise. They first present an overarching view that links three elements: the evaluation, the claim, and the consumer of that claim. They categorize common pitfalls as "sins" that occur when there is a misalignment between these elements.
They then focus on sins of reasoning, which they define as mismatches between the scope of the evaluation and the scope of the claim. These are grouped into three types - sins of ignorance, inappropriateness, and inconsistency. They also describe two sins of exposition (sins of inscrutability and sins of irreproducibility), which concern how claims are communicated and can arise through omission, ambiguity, or distortion.
Finally, the authors call for a cultural shift within research communities. They argue that work with strong evaluation but low novelty, and work with high novelty but weaker evaluation, should still be valued contributions, as both provide useful insights when claims are sound and clearly communicated.

Positive Points

- The use of figures is very effective and helps illustrate the framework visually, especially Figures 2, 9, and 10. Other figures are tied to specific examples and help contextualize them.
- The paper provides actionable guidance through its reflective questions in section 6. This makes it easier for researchers to avoid common pitfalls in their own work.
- The paper encourages a cultural shift towards valuing sound claims and rigorous evaluation, even when novelty is low. It is rare to see a paper dare to try and challenge the culture of paper selecting.

Negative Points

- Some examples are quite domain-specific, which may make it difficult for readers from other fields to fully grasp their context or the reasoning errors being illustrated.
- The paper does not offer guidance on how to prioritize or weigh the different types of sins in terms of their severity or impact.

Future Work

For future work, it would be interesting to conduct a study analyzing how these sins are distributed across published papers. We could select papers from different venues or fields, and manually classify them based on the framework. This could reveal which sins are most common in each area and provide insight into field-specific patterns - for example, in HCI papers, where involving human participants can introduce different kinds of evaluation challenges. Such a study could help raise awareness about common pitfalls within each community and potentially guide better practices in both publishing and reviewing.

Rating

I give this paper a 4/5. It gave me a clearer understanding of the relationship between research claims and their evaluation, and it will influence how I approach my own work.

Discussion Points

- The framework was developed in the context of programming languages and systems. How well do these "sins" map to other subfields (e.g., machine learning, HCI, software engineering)? Are there field-specific issues it misses?

- Should some sins be considered more severe than others? How might we prioritize or weight their impact on research quality?
- Could automated tools or reviewer aids be developed to detect ambiguous claims, missing data, or unfair comparisons?

Class Discussion Summary

- **Mapping the sins framework to other fields:**
  The discussion included perspectives from various areas of computer science, such as programming languages, machine learning, and AI. The sin of inconsistency was seen as relevant in programming languages, where comparing different languages with different source code and functionality can be unfair, and similarly in AI when models are compared solely based on performance. Sins of exposition were noted to be common in machine learning and AI due to limited data availability and the tendency to draw claims based on incomplete data. Overall, the class felt the framework is general enough to be applied to fields beyond those discussed in the paper.
- **Ranking of the sins:**
  Some participants argued that the sin of inconsistency is the most severe, as drawing conclusions from "apples-to-oranges" comparisons can be especially harmful to the community. Others felt that sins of exposition are less forgivable because they are more easily avoidable. Inadequate communication of claims or evaluations can also mislead the audience into assuming other reasoning-related sins are present.
- **Tools to detect unsound claims:**
  AI-based tools were suggested as a possible aid but were considered unreliable due to issues like hallucinations. A rule-based approach was also proposed, though participants acknowledged it would be challenging to design and maintain.