Recommendations for the Development and Assessment of AI Systems

Robert R. Hoffman, Ph.D.
[rhoffman@ihmc.us]

Institute for Human and Machine Cognition

27th International Command and Control Research Symposium

October, 2022

Acknowledgements

Timothy MILLER
University of Melbourne

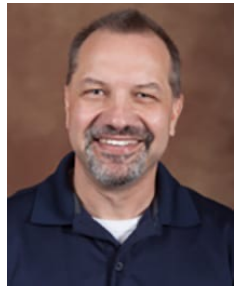
Mohammadreza JALAEIN
Macrocognition, LLC

Gary KLEIN
Macrocognition, LLC

William J. CLANCEY
IHMC

Shane T. MUELLER
Michigan Technological University

Florian JENTSCH
University of Central Florida



This research was developed with funding from
the Defense Advanced Research Projects Agency (DARPA)

Background and Context

- Experimental psychologists
  - Specialize in experimental design; cognitive systems engineering

- Experience with evaluations of computer-mediated work systems
  - Focus on human-AI performance
  - Focus on human reasoning and decision making

AI Systems Bring New and Unique Challenges

- The Human-AI work system depends critically on the cognitive capacities of both
- The context is complex and dynamic
- The AI is a “black box”
- How can a warfighter rely upon AI with confidence?

Procurement Challenges

The development of AI systems represents a significant investment.

Empirical evaluation is necessary to realize the promise of that investment

The evaluation has to examine the understandability, usability and usefulness of the technology

There must be a demonstration of the value that the technology adds to performance in the actual work context

What's the Immediate Problem?

Web search for "explanation" and "explainability" 2018 and 2021 (the duration of the DARPA XAI Program).

Only 28 reports included examples of actual machine-generated explanations.

Only 6 reported results from the evaluation of the performance of the Human-XAI duo, using such measures as decision correctness and decision time, and having type of explanation as an independent variable in the study design.

What's the Immediate Problem?

Failure to include descriptions of the actual task instructions presented to the participants.

Inadequate description of the experiment procedure.

Burdening the participants with multiple measurements and scale judgments.

Confusing description of interfaces.

Weak types of explanations are often utilized (e.g., histograph of likelihoods of outcomes)

What's the Immediate Problem?

Confusing and insufficient description of the conditions and independent variables

Lack of appropriate controls

Confounding of variables

Sample sizes that are either way too small or way too large

Confusing and insufficient presentation of results

Frequent reliance on a notion of "marginal significance"

Unnecessarily complex and confusing statistical analyses

A Major Problem is "Over-design"

Unnecessarily long sessions or trials

Needlessly large pool of participants

Too many experimental conditions and comparisons

Attempt to test too many hypotheses in a single experiment.

Too many dependent variables

Looking for effects that should manifest as three-way interactions.

Recommendations

Recommendation 1
Prior to project start, conduct an evaluation of the cognitive requirements for the envisioned work system

Recommendation 2
The team of researchers who are designing and conducting the evaluations of the AI system should be interdisciplinary.

Recommendation 3
Understand the difference between evaluation studies laboratory experiments

Recommendations

Recommendation 4

Conduct a Premortem

Recommendation 5

The group of participants in evaluation studies should include individuals who are practitioners in the domain

Recommendation 6

The evaluation should represent the actual work context

Recommendation 7

Conduct small-scale studies that are targeted to particular hypotheses

Recommendations

Recommendation 8

Displays or charts of the results on an evaluation should follow conventions provided by the American Psychological Association and the Human Factors and Ergonomics Society

Recommendation 9

Regard statistical tests as exploratory tools

Recommendation 10

Set a high bar for determining the practical value of the AI

(A Few More) Recommendations

Stop calling the evaluation studies “experiments.”

Stop calling the research participants "subjects."

Present participants clear instructions on the nature of the task,

the rationale for the task, the materials, and the AI

Avoid trying to measure everything

The data graphs presented by all Performer Teams on a project should be

consistent in their graphing design, its style and format

https://www.ihmc.us/technical-reports-on-explainable-ai/

XAI Literature Review

☞ The “Stakeholder Playbook”

XAI Metrics

Measurement of user “Mental Models”

Measuring Trust in AI systems

☞ Requirements for AI Assessment
(Minimum Necessary Rigor)

Thank you!

Questions . . . .