Proceedings of the 27th International Command and Control Research and Technology Symposium (ICCRTS 27). International Command and Control Institute, 2022.

Submission 060, Topic 5

Top Ten Recommendations for the Development and Assessment of AI Systems

Robert R. Hoffman, Ph.D.[a]
Mohammadreza Jalaeian, Ph.D.[b]
Gary Klein, Ph.D.[b]
Florian Jentsch, Ph.D.[c]
William J. Clancey, Ph.D.[a]
Shane T. Mueller, Ph.D.[d]

Introduction

Artificial Intelligence (AI) systems bring with them new and unique challenges. How can a warfighter trust a "black box"? How can a warfighter rely upon it with confidence? The development of AI systems represents a significant investment. Empirical evaluation is necessary to realize the promise of that investment by answering important questions.

Automation is a capability of a larger work system that is enabled by the integration of human and machine abilities.[1] The Human-AI work system depends critically on the cognitive capacities of both partners, in a context that is complex and dynamic. The evaluation of AI systems has to be an unbiased examination of the understandability, usability and usefulness of the technology, and a demonstration of the value that the technology adds to performance in the actual work context.[2]

A survey of software procurement activities[3] revealed that trillions of dollars have been wasted in failed procurements, because of shortfalls in system usability and usefulness. Shortfalls have direct impact on the people who rely on the systems. The shortfalls might be remedied by certain changes in policy, and that is the focus of this presentation. The recommendations derive from the experience of the authors—cognitive systems engineers who have been involved in AI system development and evaluation for decades. The recommendations should be applicable to all research intended to evaluate the effectiveness of AI systems.

Recommendations

In the creation of AI systems, computer scientists commence the development process even during the program proposal phase. This typically proceeds without the consideration of the AI as one component in a Human-AI work system. Notional computational architectures are envisioned,

[a] Florida Institute for Human and Machine Cognition
[b.] Macrocognition LLC
[c.] University of Central Florida
[d.] Michigan Technological University

ones that are hypothesized to be promising solutions. But because the developers "jump in," they may be basing their envisioned technologies on limited notions of the cognitive work.

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

Even before a Program is announced, an advisory group, especially cognitive systems engineers, should recommend to Proposers the cognitive requirements that the Human-AI work system should satisfy. Cognitive work analysis[4] reveals what the operator needs from the technology, which is often contrary to what the system developer assumes.

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

The team should preferably include: (1) expert operators who are accomplished in the target domain, (2) specialists in in human-computer interaction, computer-supported cooperative work, or human factors, (3) cognitive/social scientists, and (4) specialists in psychological research methodology and psychometrics.

The best way to ensure usability and usefulness of an AI program is to engage domain experts during the R&D process. Experience with large-scale projects (e.g., aircraft, highly automated naval vessels, spacecraft, self-driving vehicles) and their ensuing failures indicates that the evaluation of automation is crucial early in design, *before* iterative prototyping[2]. Having engaged a cohort of practitioners in guiding the technology development process from the very start, one would have by the completion of the evaluation a cohort who are facile in using the new technology[5]. They would be the best people to begin training others, and training the trainers.

Recommendation 3: *Understand the difference between evaluation studies laboratory experiments.*

Laboratory experiments require rigor that is not necessary for evaluations of prototypes in work settings (not to be confused with “demonstrations” that use a rehearsed script). A phrase such as "empirical evaluation" might be preferred. In particular, evaluation should involve meaningful test trials with informed operators in authentic work settings. Only by doing that can results be useful for refining the new work system. It is important to avoid the slippery slope of complex designs with multiple factors and control conditions, and drawn-out series of experiments.

As a case in point, a federally funded research and development center recently held a meeting to review a new technology procurement. At that meeting a participant stated that the program would need to include experimentation. A senior government official, interpreting this to entail protocols like those of laboratory experiments, responded that the military no longer seemed very enthusiastic, because of many experiences in which the research was too expensive, took too long, and provided data that were obsolete by the time they arrived.

Recommendation 4: *Conduct a Premortem.*

A Premortem is a proven process in which the members of the development team generate and discuss reasons why the project might not work.[6] Reasons can span any aspect of the project, from system architecture, to interfaces, and to the evaluation process itself. The reasons are collectively assessed to generate concepts of how to anticipate and avoid or mitigate traps or shortcomings. The Premortem activity can take less than an hour.

Recommendation 5: *The group of participants in evaluation studies should include individuals who are practitioners in the domain.*

Too often, AI assessment only involves "Mechanical Turkers" as the research participants. Certainly, there is a need to see to see what happens when novices first learn the tasks, so some of the participants can be college students or some slice of a general population. But it is absolutely necessary for evaluation to include experienced operators. A root cause of software system shortfalls is the failure to involve the operators (or end-users) in system development and evaluation.[3]

Recommendation 6: *The evaluation should represent the actual work context.*

In the experimental laboratory, the tasks presented to participants are often simplified versions of real-life tasks or experiences. They often end up artificial, so-called "toy" problems. They are abstracted away from their context in service of experimental control.

In many AI evaluation projects, system developers put themselves in the shoes of the end-user, and assume what the end-user needs. But the tasks, materials, interface, etc. should all be representative of the tasks that are conducted in the work domain and setting for which the AI has been created. "It is important that operational experience is communicated to the development community so that lessons in the field can ultimately influence upgrades to existing systems and the designs of future systems."[7]

In evaluations of prototype AI systems, there is pressure to resort to artificial tasks because they are easier to design and conduct. But the tasks should preserve the context and complexities of the operational environment[2]. This is often called a "demonstration" in DoD contexts. These activities do not adhere to rehearsed scripts, but allow for free-form operations in meaningful situations involving multiple people interacting with multiple technologies. When developers of AI systems evaluate AI prototypes with de-contextualized artificial tasks, the multi-faceted character and variability of the work context is lost. As a consequence, the results from such sterilized evaluations may not carry over to actual operations.

Recommendation 7: *Conduct small-scale studies that are targeted to particular hypotheses.*

In contrast with the demonstrations that are used to evaluate the readiness of technology, in the evaluation of AI systems the developers typically conduct single, large-scale, complex laboratory-like experiments. They attempt to test many hypotheses all at once. This requires many conditions and controls. Large samples are required to reach the desired statistical effect sizes. Evaluation sometimes has hundreds of people who engage with the AI program using such platforms as Mechanical Turk. Rather than undertaking such an effort, small-scale studies with limited hypotheses are advised—if there is no clear gain from a technology intervention with a sample of 10 participants, then you know something is very wrong with the system design.

Recommendation 8: *Displays or charts of the results on an evaluation should follow conventions provided by the American Psychological Association[8] and the Human Factors and Ergonomics Society[9].*

Very few reports evaluating AI systems portray results by well-designed graphs or charts. Many of the problems seen in data graphs derive from researchers relying on readily-available software. The graph defaults impose a style that is violates best practices in human factors and experimental psychology. Graphing improprieties take a number of forms: the use of microfonts, the use of thin lines for the graph axes, the use of under-specific labels for the axes, too much

clutter, over-use of color coding, and so forth. To the even greater detriment of system development, we observe that different performer teams within a single funded research program will use different graphing styles. Conformance to accepted standards on the part of all the teams would facilitate comparing results and enhance the clarity of communication among the researchers and for those who might build upon their work.

Recommendation 9*: Regard statistical tests as exploratory tools.*

In addition to falling into traps designing evaluations and graphically presenting the results, developers of AI systems fall prey to statistical traps, including misuses and misunderstandings of traditional significance testing.[10,11] The quest for statistical significance and (even modest) effect sizes is treated as definitive and final proof of a causal hypothesis. Developers often report on statistical analyses that are opaque and complicated, here too assuming that laboratory experimentation should be the template for evaluating redesigned work systems.

Statistical significance testing is an exploratory tool, not a means for automating scientific judgment by calculating significance levels or effect sizes. The reliance on significance testing is important and necessary, but should be tempered by considering the practical significance of the results. The performance of the work system should show marked improvement.[12] Of course, the scale of "significance" depends on the domain and the nature of the work. In some domains, even saving one minute by a single operator can be of immense practical significance. The determination of practical significance has to be a matter for expert judgment as well as statistical calculation.

Recommendation 10: *Set a high bar for determining the practical value of the AI program.*

Interviews with operators and stakeholders have revealed that they often set a very high bar in the operational setting. In an interview with stakeholders, one of them said that if he could not achieve an understanding of how an AI system works within ten trials or attempts, that he simply would not use it.[13] Another said that unless a new tool enabled successful performance on 85 percent of the key tasks on the very first use, then the tool would not be welcome.

In the field setting, operators may have to use an AI system with minimum training, entailing a demand that AI systems be highly learnable, if not intuitive. A study by David Klinger and his associates[14] involved the re-design of a workstation and its interfaces for operators on the AWACS air defense platform. A task analysis revealed 40 problems with the existing interface that made the work inefficient (e.g., poorly designed displays, unnecessary memory demands, loss of situational awareness). The results suggested a redesign, which was implemented and then evaluated. But the opportunity for the operators to learn and then perform with the new workstation was very limited, to only about five hours. In contrast, the participants had had hundreds of hours of practice with the existing interface. Yet their performance with the new interface showed a notable improvement relative to baseline performance. This was a very simple study design: One experimental condition (the new interface) compared to archived data on baseline performance, and a relatively small sample size (18 operators).

Conclusion

Experience with AI assessment has revealed limitations and inefficiencies in the standard practices used to evaluate technology designed for operational settings. The recommendations presented here are intended as contributions to the development of AI measurement science, broadly, to help ensure that AI systems are understandable, usable, useful, safe, and trustworthy.

Acknowledgement and Disclaimer

This material is approved for public release. Distribution is unlimited. This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

End Notes

1. Defense Science Board (2012). "The Role of Autonomy in DoD systems." Washington, DC: Department of Defense.

2. William J. Clancey (2020). *Designing agents for people: Case studies of the Brahms work practice simulation framework.* Amazon Kindle Print Replica e-book.

[https://www.amazon.com/Designing-Agents-People-Simulation-Framework-ebook/dp/B08D7XK8ZY].

3. The Standish Group (2014). "The Standish Group Report: Chaos." Project Smart, The Standish Group International. [https://www.standishgroup.com]

4. Beth Crandall, Gary Klein, and Robert R. Hoffman (2006). *Working Minds: A Practitioner's guide to cognitive task analysis.* Cambridge, MA: MIT Press.

5. Steven V. Deal and Robert R. Hoffman (2010, March/April). The Practitioner's Cycles, Part 1: The Actual World Problem. *IEEE Intelligent Systems*, pp. 4-9.

6. Gary Klein (2007). Performing a project premortem. Harvard Business Review, 85(9), 18-19.

7. Defense Science Board (2012). p.2.

8. American Psychological Association (2022). Graphing Guidelines [https://apastyle.apa.org/style-grammar-guidelines/tables-figures/figures]

9. Human Factors and Ergonomics Society (2022). Guidelines for Presenting Quantitative Data. [https://www.researchgate.net/publication/220457627_Guidelines_for_Presenting_Quantitative_Data_in_HFES_Publications]

10. Gerd Gigerenzer (2004). Mindless statistics. *Journal of Socio-Economics, 33,* 587–606.

11. Robert R. Hoffman (2020, September). Concept Blog Episode No. 5: "0.01 and 0.05." [https://www.ihmc.us/hoffmans-concept-blog/]

12. Robert R. Hoffman, Gary Klein, Shane T. Mueller, and William J. Clancey (2021). "Recommendations for the Empirical Assessment of Human-AI Work Systems: A Contribution to AI Measurement Science." Technical Report from Task Area 2 to the DARPA Explainable AI Program. [https://www.ihmc.us/technical-reports-on-explainable-ai/]

13. Robert R. Hoffman, Gary Klein, Mohammadreza Jalaein, Shane T. Mueller, and Connor Tate (2021). "The Stakeholder Playbook." Technical Report, DARPA Explainable AI Program. [https://psyarxiv.com/9pqez/]

14. David W. Klinger, S. J. Andriole, L.G. Militello, Leonard Adelman and Gary Klein (1993). “Designing for performance: A cognitive systems engineering approach to modifying an AWACS Human -Computer Interface.” Technical Report AL/CF-TR-10993-0093.Armstrong Laboratory, Air Force Materiel Command, Wright-Patterson Air Force Base, OH.