Received: 13 February 2022 | Revised: 14 March 2023 | Accepted: 14 July 2023

DOI: 10.1002/aaai.12108

ARTICLE

"Minimum Necessary Rigor" in empirically evaluating human–AI work systems

Gary Klein[1] | **Robert R. Hoffman**[2] | **William J. Clancey**[2] | **Shane T. Mueller**[3] | **Florian Jentsch**[4] | **Mohammadreza Jalaeian**[5]

[1]MacroCognition, LLC, Washington, District of Columbia, USA

[2]Institute for Human and Machine Cognition, Pensacola, Florida, USA

[3]Michigan Technological University, Houghton, Michigan, USA

[4]University of Central Florida, Orlando, Florida, USA

[5]The Ohio State University, Columbus, Ohio, USA

Correspondence
Robert R. Hoffman, Institute for Human and Machine Cognition, Pensacola, Florida, USA.
Email: rhoffman@ihmc.us

Funding information
Defense Advanced Research Projects Agency

Abstract

The development of AI systems represents a significant investment of funds and time. Assessment is necessary in order to determine whether that investment has paid off. Empirical evaluation of systems in which humans and AI systems act interdependently to accomplish tasks must provide convincing empirical evidence that the work system is learnable and that the technology is usable and useful. We argue that the assessment of human–AI (HAI) systems must be effective but must also be efficient. Bench testing of a prototype of an HAI system cannot require extensive series of large-scale experiments with complex designs. Some of the constraints that are imposed in traditional laboratory research just are not appropriate for the empirical evaluation of HAI systems. We present requirements for avoiding "unnecessary rigor." They cover study design, research methods, statistical analyses, and online experimentation. These should be applicable to all research intended to evaluate the effectiveness of HAI systems.

INTRODUCTION

The development of AI systems represents a significant investment, and empirical testing is necessary in order to realize the promise of that investment. This article considers the empirical evaluation of human–AI (HAI) work systems. HAI systems are ones in which humans work in an interdependence relationship with AI tools in order to conduct work-related tasks (Clancey and Euchner 2021; Johnson and Vera, 2019). HAI systems are formative in such domains as emergency response management, industrial process control, health care, business, banking and finance, military command and control, autonomous vehicles, transportation systems, weather forecasting, and so forth.

The empirical evaluation of AI has a considerable foundation. Cohen and Howe (1988) and Cohen (1991, 1995) focused on methods for evaluating claims made about the performance of programs, especially exploratory and statistical data analysis. Hoffman (1992) focused on the application of psychological methods for efficient knowledge capture in the development of knowledge bases for expert systems. Nielsen (1997) discussed the empirical assessment of usability, that is, user testing of web pages. None of these involved the empirical assessment of the performance of HAI systems. Hernandez-Orallo (2017a,b) reviewed AI application areas and evaluation approaches (peer reviews, competitions, achievement of performance thresholds, measures of efficiency). The focus was on mathematical methods for comparing AI and human

cognitive competency. Only a few of these approaches involved empirical study of the performance of a HAI work system.

Recent research has begun to address this matter of experimental adequacy and rigor in empirical evaluation of performance. Researchers have argued that evaluative experimentation has to involve tasks that are pertinent to actual applications "using tasks, data, users, and metrics grounded in the intended deployment contexts" (Amarasinghe et al. (2022, p. 1), rather than relying on simplified proxy tasks (Buçinca et al. 2020). In the rapidly developing field of explainable AI, Lage et al. (2019) advocate for carefully controlled human subjects experimentation to determine whether machine-generated explanations actually have explanatory value to users and have a positive impact on performance. The expansive literatures that are pertinent to explainable AI point to a number of challenges with regard to evaluation, measurement, and metrics. These are detailed in Hoffman et al. (2023) and Mueller et al. (2019, 2020). The need for careful consideration of experimental design is highlighted by the finding of Amarasinghe et al. (2022) that "seemingly trivial experimental design choices can yield misleading conclusions" (p. 1). The requirements that are proposed the present article reflect and expand these concerns about methodology.

In the disciplines of procurement, computer science, and psychology, the word "experiment" carries quite different meanings. In systems acquisition, what are often called experiments are what experimentalists would call demonstrations. Although not "experiments" is a laboratory sense, empirical assessment occurs at every step in the system development process, spanning requirements development, system design, implementation, validation and verification testing, and refinement phases. The word "evaluation" in the present article is not used in this comprehensive sense. Instead, this article focuses on the empirical evaluation of the performance of the HAI work system, using human research participants to demonstrate the value of the AI system, or a prototype of an AI system.

The requirements that are proposed here refer to the methodology of HAI assessment. The guidance derives from the collective experience of the authors, who have been involved in AI system development and evaluation for decades, witnessing the problems that result when considerations that are basic to cognitive systems engineering and experimental psychology are not fully appreciated and adopted. The guidance presented in this article is intended primarily for system developers. The goal is to promote meaningful trans-disciplinary research that meets the needs of sponsors and the beneficiaries of AI tools, and that encourages sponsors to support rather than avoid research.

This article is organized as follows. The first two sections set the stage by specifying what experimentation must accomplish and what experimental "rigor" means in the context of AI Measurement Science. The third section presents the methodological requirements and recommendations.

WHAT MUST AI ASSESSMENT ACCOMPLISH?

The HAI system depends critically on the capacities of both the human and the AI, working in a context that is complex and dynamic. An empirical investigation that is intended to assess the quality of the work is essentially a psychological experiment, one in which the equipment with which the research participants work is an expensive computational system. Performance evaluation must demonstrate that the work method that is shaped by the AI is understandable, learnable, usable, and useful. When considered from this perspective, a host of questions about the evaluation confront the developer: How do we distinguish the human and AI contributions to the HAI system? How do we test for the usefulness and usability of the AI within the HAI system? Of the human? How do we evaluate work system performance? Is the AI valuable to users in their actual work context? What are our measurement scales and metrics? Is the work process that is imposed by the AI one that can be readily learned?

In addition to assessment, empirical evaluation must be a path to discoveries. Researchers should be open to surprises, and be prepared to exploit what is learned. When considered from this perspective, a host of additional questions challenge the developer, such as: Does the AI enable the user to diagnose AI limitations, edge cases, and difficult situations? Does the AI empower the user to create kludges and work-arounds? Does the AI enable the user to learn about what can go wrong? Does the AI empower the user to recover from mistakes? Do the task and the AI enable the participant to increase their expertise?

WHAT IS RIGOR IN AI MEASUREMENT SCIENCE?

In laboratory experimentation, control and manipulation of variables is necessary. This invariably entails a need to conduct multiple experiments. In the pragmatic context of the evaluation of AI systems, the studies need to be efficient. Yet, many variables play a role in determining the AI-enabled work and its outcome. Whether the work is simply being observed or the work is being manipulated (i.e., by the insertion of new technology), multiple

variables need to be taken into account. This leads to the "fundamental disconnect": The time frame for effective experimentation––with "effective" defined in terms of the strictures of the experimental laboratory––is outpaced by the time frame for change in the work and its technology (Hoffman et al. 2010). It is desirable to avoid the problem of multi-year evaluation because the technology is likely to be substantially modified even while the evaluation is taking place. From the perspective of technology development, developers do not want to wait while numerous complex experiments are conducted.

In its word origins, "rigor" means inflexibility. The word "experiment" carries with it the assumption that the empirical activity must be tightly controlled, as in a laboratory. Researchers can more add rigor to toy problems and laboratory-like tasks than to realistic tasks. But more rigor, as defined in laboratory experimentation, is not necessarily better when transposed to the context of evaluating HAI systems. It can even be worse. The imposition of the rigor requirements of the controlled laboratory instills a tendency to over-control variables and decontextualize the tasks, making them more artificial. Stripping the context away may lead to findings that do not apply to the sponsor's needs. Thus, unnecessary rigor can create barriers to evaluation. We call this "rigor mortis."

Rigor mortis often sets in at the very beginning of a research project when there is a programmatic requirement to produce a literature review as an early deliverable. Literature reviews always seem obligatory, but are rarely created quickly enough to actually impact the technology development, which proceeds apace, and at risk. Best practice is to identify the traps and challenges discovered in previous work on the topic at hand, and do so even *before* the program description is cast in stone. The direct path to that would be interviews with a small number of selected leaders or experts in the pertinent fields. Those individuals would provide the most succinct and important historical scholarship. That should be obtained prior to the inception of an AI research and development project.

Rigor mortis also emerges as projects proceed, as the following case study shows. The research team tried to take on too much and measure too many things.

Rigor mortis experiences have discouraged many government sponsors from conducting any evaluations at all.

It has been noted within the Department of Defense that there is a need to develop a research roadmap for implementing AI (Government Accountability Office 2022; Trent and Doty 2022). At a minimum, it is necessary to demonstrate that the AI technology results in an improvement in the performance of the work system. The

Case study in rigor mortis

A government agency funded a large-scale study to compare conventional aircraft cockpits, with electro-mechanical instruments, to new "glass cockpits" (digital instrument displays) to see what the new technology contributed and what its limitations were. A large team of contractors and government researchers were involved in all this work––it was going to be a landmark project, a career-defining set of experiments to serve as a standard for doing good science on an applied question about a human–machine work system. The study involved carefully controlled conditions, carefully selected scenarios, and large numbers of commercial pilots to be research participants. It took a year just to design this single, complex experiment. Data were collected on a large number of variables, to make sure little got missed, rather than to evaluate targeted hypotheses. It took another year to run all the participants. Then came the challenge of how to analyze all the data, and it took another year to develop the evaluation plan. These years of delay made the results less relevant than they could have been years earlier, plus entailing such high levels of complexity for the data analysis that no one was willing to step in when the government project monitor transitioned to another program. The project was terminated. The data were never analyzed.

Case study in the consequences of rigor mortis

Recently, a federally funded research and development center held a meeting to review a new technology procurement. At that meeting, a participant stated that the program would need to include a performance evaluation. A senior government official responded that the military no longer seemed very enthusiastic about research and experiments, because of many experiences where the research was too expensive, took too long, and provided answers that were obsolete by the time they arrived.

Case study in minimum necessary rigor: The "Klinger–Klein test"

A study by Klinger et al. (1993) illustrates the practical constraints that can be involved in the evaluation context, and how it is possible to satisfy the "lightweight yet necessary" requirement despite those constraints. The project involved the design of a workstation and its interfaces for Weapons Directors on the Airborne Warning And Control System, an air defense platform. A cognitive task analysis revealed 40 problems with the existing interface that made the cognitive 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 Weapons Directors to learn and then perform with the new workstation was very limited, to only four and a half hours. The Weapons Directors 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 experimental design: One experimental condition (the new interface) compared to a control condition (archived baseline performance data), and a relatively small sample size (18 Weapons Directors). Many features of the redesigned workstation were incorporated in the next evolution of the AWACS system.

following case study illustrates what it means for an evaluation to be *sufficient*: it was a simple experimental design that demonstrated the value added.

The next sections present the requirements for Minimum Necessary Rigor (MNR) in AI evaluation methodology. Their objective of these requirements is to reduce or eliminate the excessive expense and excessive time.

THE PARTICIPANTS

Requirement 1: *Individuals who would be the beneficiaries of the proposed AI system (that is, domain practitioners or operators) should be involved throughout AI system design and development, including involvement as actual research participants.*

AI evaluations seem to rarely use domain experts or system operators as research participants. Instead, they rely on novices or on samples of convenience. That said, for some tasks and applications there may be a need to see what happens when novices first learn the tasks. The issue of participant selection can get complicated if the AI system is designed for a variety of types of users, ranging in experience. Nevertheless, if an AI system is designed to help domain practitioners, it should use them in the evaluation (Deal and Hoffman 2010).

Requirement 2: *Training of the research participants should be minimal.*

In the field setting, users may have to rely on an AI system with minimum training. Therefore, AI systems should be readily learnable, if not intuitive. Researchers may want to evaluate a few groups receiving different types or amounts of training, but this is not necessary, as long as satisfactory performance can be achieved following minimal training.

Requirement 3: *The number of research participants in the study conditions does not have to be large.*

Researchers usually desire experiments with large-n. They know that by increasing the sample size, they increase their chances of achieving statistical significance on a parametric test. Further, it is now easy to collect large data sets via online platforms. So why not have a large-n if it is easy to get? This is a mythical belief. Large data sets from complex factorial experiments mandate significant efforts at data analysis, and the explanation of the results gets convoluted. The trade-off is that the relatively lower effort to get the data is balanced by the relatively greater effort to make sense of the data.

Psychological research seeks effects that obtain for a majority of the research participants, even in small samples. For example, in the evaluation of an automated decision aid, one would want (or need) to see an effect (i.e., the decision aid helps) with samples as small as 10 participants. If the sample size in any one condition is less than about 10, there is considerable risk of confusing individual differences with main effects, especially if there is some bias in the selection of the participants, or a task demand that influences performance. But the point of Requirement 3 is this: If you cannot get a clear effect of a technological intervention on a sample of 10 participants, then something is wrong.

According to the common parametric methods, one can increase confidence that a discovered difference is "real" (the power of a statistical test) by increasing the sample size, but that makes the test sensitive to smaller differences. For a 1.0 standard deviation difference between the means of experimental and control groups (considered a large effect size), there may be over 80% overlap of the groups' frequency distributions (Sullivan and Fein 2012). To concretize this, increasing the sample size from 10 (say) to 100 may lend greater confidence, but that might be

confidence about a difference that is too small to really matter.

It is also important to note that one can obtain a richer and clearer idea of what is going on in participant reasoning by in-depth cognitive interviews with five to seven participants, rather than by running a large-n experiment with a fixed-response questionnaire tacked on at the end (Crispen and Hoffman 2016).

THE TASKS

Requirement 4: *The participant's task should be ecologically valid*.

In too many evaluations, the task presented to participants is only tenuously pertinent to the intended application domain. Evaluation research often resorts to artificial tasks because they are easiest to design and present. An example would be to use the task of identifying different sports activities in photographs when the ultimate application would be to recognize suspicious activities. The task is removed from its "real world" context. In the empirical assessment of HAI systems, the task presented to participants should be one that is representative of the tasks that are conducted in the work domain for which the AI has been created (see Buçinca et al. 2020; Clancey 2020; Clancey et al. 2011; Hernandez-Orallo 2017b). If the evaluation is sterilized, the results may be irrelevant. Ideally, evaluators can try out their AI system during actual work, or perhaps in a training exercise, or failing that, in a simulation. Also, it is well known that the cognitive requirements of the task can matter more than the surface features, so researchers need not worry about detailed replication of the look-and-feel of the actual task (so-called physical fidelity)––but should focus more on capturing the things that make the actual task cognitively difficult for the intended users.

PREPARING THE EXPERIMENT

Requirement 5: *Target particular hypotheses*.

All too often, researchers compose single, very complex experiments on the assumption that large-scale experiments—involving multiple conditions, large-n, and the manipulation and measurement of multiple variables—can adequately evaluate multiple hypotheses. Unfortunately, such studies require compromises, and typically end up achieving none of their goals; they do not answer one question well, nor do they answer many questions sufficiently.

Requirement 6: *Conduct pilot studies to test and refine the methods, materials, and procedures*.

Too many evaluation studies dive into large-scale experiments, and once started, adjustment of the method and procedure always causes complications. Best practice is to conduct one or more pilot studies (Cohen 1995). These are *not* designed to evaluate the primary hypothesis (e.g., whether the technology intervention is good), but instead are intended to garner assurance that the methodology is sound and the procedure runs smoothly. Almost invariably, pilot studies lead to improvements in the study design and methods or the procedural details of conducting the evaluation. Pilot studies can involve as few as 10 research participants.

DESIGNING THE EXPERIMENT

Requirement 7: *Run a two-condition, between-participants study*.

Consider two conditions, which we call Evaluation and Control. Different individuals would participate in the two conditions. The Evaluation condition would involve the AI, the Control condition would not. Research participants in both conditions would perform the same task. This assumes, of course, that the task as completed in the Evaluation condition is the same as the task that is used in the Control condition. The purpose of this study is to demonstrate that the technology insertion is good.

Evaluation condition	Control condition
Participants 1 through n	Participants $n + 1$ through $2n$

An alternative design is to have a Control condition in which the new technology is inserted, but some crucial element or capability of the new technology is disabled. A number of the key elements of the technology might be hobbled altogether. Assuming that the tasks in the two conditions are equated for difficulty, if the results do not clearly distinguish the Control and Evaluation conditions, something is wrong. If the results do clearly distinguish the Control and Evaluation conditions, subsequent studies can engage in more targeted hypothesis tests. Note that in this design it may not be necessary to have a Control condition if there are usable baseline data on performance using the legacy work system.

Requirement 8: *Run a two-condition, within-participants study*.

A second study design also involves two conditions, which we again call Control and Evaluation, but now in the within-participants study, each participant experiences both conditions. This is a repeat-measures design. For example, in the Evaluation condition, the participants would work using a new AI decision aid, whereas in

the Control condition, the same participants would work only with the benefit of the legacy decision aid. Over trials in the Evaluation condition, performance should improve, but when the Control condition is instituted, performance should drop. If the order of the conditions is reversed, then performance should improve when participants move from the Control to the Evaluation condition. However, the experiment need not involve both orders (counterbalancing). One is order enough to satisfy MNR.

Condition 1: Evaluation	Condition 2: Control
Participants 1 through *n*	Participants 1 through *n*

The Evaluation conditions in the between-participants design as well as the within-participants design have the benefit of permitting an investigation of the learning curve for using the AI technology. The early trials in the Evaluation conditions are, effectively, training. As participants engage in the work over a series of trials or test cases, their performance should improve. The form of the learning curve can help you project the scope that will be required of a training regimen.

ANALYZING THE RESULTS

Requirement 9: *Do not apply statistical analyses that are opaque or complicated*.

If the AI does not yield a dramatic improvement in performance, why go to the trouble of developing it and training people to use it in the field? Especially unnecessary is the concern over achieving statistical significance at the $p < 0.01$ level versus the $p < 0.05$ level, or obtaining results that are described as "nearly" or "marginally" significant, with the probability values reported out to the fifth decimal place. On some interpretations of statistical significance, the decision is binary and therefore wishful thinking is not a legitimate basis for making decisions about a null hypothesis (see Hoffman, 2020). Also, the sponsors of the research are unlikely to understand or appreciate the statistical gyrations taken to tease statistically significant results from minimal effects.

Requirement 10: *Be prepared to set a high bar for determining whether or not the AI is good*.

Interviews with users and stakeholders have revealed the "high bar" that is set in the field 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 10 trials or attempts, then he simply would not use it (Hoffman et al. 2021). Another said that unless a new tool enabled successful performance on 85% of the key tasks on first use, then the tool would not be desired.

Requirement 11: *Consider practical significance*.

It is common for researchers, especially in applied contexts, to see the achievement of statistical significance as a key metric or decision point, without considering practical significance. In an inferential leap, statistical significance is taken as proof of a causal relation between the dependent and independent variables, even when the difference between group means is small and the distributions overlap considerably. Statistical significance is a qualified aspect of importance—results can be statistically significant even if they are not useful or do not have practical significance.

The notion of practical significance has a considerable history. In 1919, the famed psychologist Edwin Boring referred to the difference between mathematical and "scientific significance." In 1954, Hodges and Lehmann assessed the validity of statistical hypotheses in a comparison of statistical significance and "material significance." In 1956, Roger Kirk introduced the phrase "practical significance," saying:

> *"The appeal of null hypothesis significance testing is that it is considered to be an objective, scientific procedure for advancing knowledge. In fact, focusing on p- values and rejecting null hypotheses actually distance us from our real goals: deciding whether data support our scientific hypotheses and are practically significant or useful" (pp. 755).*

However, there has been no progress on developing calculational methods, related perhaps to the fact that practical significance is a matter of judgment, and cannot be determined solely on the basis of a mathematical analysis of performance data. Klein et al. (2021) present an approach that involves combining statistical analysis with expert judgment. The statistical analysis determines whether an evaluation result indicates potential practical significance, with regard to a threshold that can be lax (i.e., any overall performance improvement is desirable) or strict (setting a high bar, as in Requirement 10, above). If performance data show that a selected threshold has been crossed, then a small panel of experienced domain practitioners can be assembled to conduct an evaluation using a judgment scale for actual practical significance.

CONCLUSION

The requirements presented here constitute a foundation or starting point for devising an efficient evaluation. If the MNR requirements are met, but the empirical results are not promising, there is no point in conducting more

elaborate or larger-scale experiments. If the MNR requirements are met, and the empirical results do show promise, then and only then might more elaborate investigations be conducted.

Some researchers would add to the requirements that are presented above. They might want to always conduct post-experimental debriefings. They might advocate for controls for task demands (i.e., what participants do and say is influenced by their being participants in an experiment). Control for this has been lacking in AI evaluations. We have not delved further into methodological details because we wanted to maintain our focus on MNR. Our experience in technology assessment has revealed some challenges to standard laboratory-centered practices in evaluation methodology. The Minimum Necessary Rigor concept is aimed at escaping the "fundamental disconnect" and improving the practice of evaluation. We hope that the requirements presented here will serve as a springboard for discussion of these matters.

ACKNOWLEDGMENTS

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. The Authors would like to thank the three anonymous Reviewers for their comments on the draft of this article.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

REFERENCES

Amarasinghe, K., K. T. Rodolfa, S. Jesus, V. Chen, V. Balayan, P. Saleiro, P. Bizarro, A. Talwalkar, and R. Ghani. 2022. "On the importance of application-grounded experimental design for evaluating explainable ML methods." [downloaded 2 August 2023 at arXiv:2206.13503].

Boring, E. G. 1919. "Mathematical vs. Scientific Significance." *Psychological Bulletin* 16: 335–8.

Buçinca, Z., P. Lin, K. Z. Gajos, and E. L. Glassman. 2020. "Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems." In *Proceedings of the ACM International Conference on Intelligent User Interfaces*, 454–64. New York: Association for Computing Machinery. downloaded 29 January 2023 from https://doi.org/10.1145/3377325.3377498

Clancey, W. J. 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

Clancey, W. J., and J. Euchner. 2021. "Scientific Design of Tools for Amplifying Intelligence." *Research-Technology Management* 64(2): 11–7. downloaded 9 March 2023 from https://doi.org/10.1080/08956308.2021.1868912

Clancey, W. J., M. Lowry, R. Nado, and M. Sierhuis. 2011. "Software Productivity of FIELD Experiments Using the Mobile Agents Open Architecture with Workflow Interoperability." In *Proceedings of the IEEE Fourth International Conference on Space Mission Challenges for Information Technology* (SMC-IT), 85–92. Palo Alto, CA: IEEE Computer Society.

Cohen, P. R. 1991. "A Survey of the Eighth National Conference on Artificial Intelligence: Pulling together or pulling apart?" *The AI Magazine* 12(1): 16–41.

Cohen, P. R. 1995. *Empirical Methods for Artificial Intelligence*. Cambridge, MA: MIT Press.

Cohen, P. R., and A. E. Howe. 1988. "How Evaluation Guides AI Research." *The AI Magazine* 9(4): 35–43.

Crispen, P., and R. R. Hoffman. 2016. "How Many Experts?" *IEEE Intelligent Systems* 31: 57–62.

Deal, S. V., and R. R. Hoffman. 2010. "The Practitioner's Cycles, Part 1: The Actual World Problem." *IEEE Intelligent Systems* 25: 4–9.

Government Accountability Office. 2022. "DOD Should Improve Strategies, Inventory Process, and Collaboration Guidance." Report GAO-22-105834, Government Accountability Office, Washington, DC. downloaded 9 March 2023 from https://www.gao.gov/products/gao-22-105834

Hernandez-Orallo, J. 2017a. "Evaluation in Artificial Intelligence: From Task-oriented to Ability-oriented Measurement." *Artificial Intelligence Review* 48: 397–447.

Hernandez-Orallo, J. 2017b. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge: Cambridge University Press.

Hodges, J. L., and E. L. Lehmann. 1954. "Testing the Approximate Validity of Statistical Hypotheses." *Journal of the Royal Statistical Society, Series B (Methodological)* 16: 261–8.

Hoffman, R.R., ed. 1992. *The Psychology of Expertise: Cognitive Research and Empirical AI*. Mahwah, NJ: Erlbaum.

Hoffman, R. R. 2020. "Concept Blog Episode No. 5: "0.01 and 0.05"." https://www.ihmc.us/hoffmans-concept-blog/

Hoffman, R. R., G. Klein, M. Jalaein, S. T. Mueller, and C. Tate. 2021. "The Stakeholder Playbook." Technical report, DARPA Explainable AI Program. https://psyarxiv.com/9pqez/

Hoffman, R. R., M. Marx, R. Amin, and P. L. McDermott. 2010. "Measurement for Evaluating the Learnability and Resilience of Methods of Cognitive Work." *Theoretical Issues in Ergonomic Science* 9(2): 213–28.

Hoffman, R. R., S. T. Mueller, G. Klein, and J. Litman. 2023. "Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance." *Frontiers in Computer Science*. downloaded 9 March 2023 from https://doi.org/10.3389/fcomp.2023.1096257/full

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

Johnson, M., and A. H. Vera. 2021, Spring. "No AI Is An Island: The Case for Teaming Intelligence." *The AI Magazine*: 17–28.

Kirk, R. E. 1996. "Practical Significance: A Concept Whose Time has Come." *Educational and Psychological Measurement* 56: 746–59.

Klein, G., M. Jalaeian, R. R. Hoffman, S. T. Mueller, and W. J. Clancey. 2021. "Requirements for the Empirical Assessment of Human-AI Work Systems." Technical report, DARPA Explainable AI Program. https://psyarxiv.com/j8t3c/

Klinger, D. W., S. J. Andriole, L. G. Militello, L. Adelman, and G. Klein. 1993. "Designing for Performance: A Cognitive Systems Engineering Approach to Modifying an AWACS Human Computer Interface." Report AL/CF-TR-1993-003, Human Engineering Division, Wright-Patterson Air Force Base, OH. Crew systems directorate, Air Force Materiel command.

Lage, I., E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez. 2019. "An Evaluation of the Human-Interpretability of Explanation." In *Proceedings of the 32nd Conference on Neural Information Processing Systems* (NIPS 2018), Montréal, Canada. downloaded 9 March 2023 at arXiv:1902.00006v2.

Mueller, S. T., R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein. 2019. "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications and Bibliography for Explainable AI." Technical Report from Task Area 2 to the DARPA Explainable AI Program [arXiv:1902.01876] https://apps.dtic.mil/sti/citations/AD1073994

Mueller, S. T., E. S. Veinott, R. R. Hoffman, G. Klein, L. Alam, T. Mamun, and W. J. Clancey. 2020. "Principles of Explanation in Human-AI Systems." In *Proceedings of the Workshop on Explainable Agency in Artificial Intelligence* (AAAI-2020). [downloaded 9 March 2023 from arXiv:2102.04972].

Sullivan, G. M., and R. Fein. 2012. "Using Effect Size––or Why the *p*-Value is Not Enough." *Journal of Medical Education* 46: 2679–282.

Trent, S., and J. Doty. 2022. "Focus on the Human Element to Win the AI Arms Race." National Interest. [downloaded 9 March 2023 from https://nationalinterest.org]

How to cite this article: Klein, Gary, Robert R. Hoffman, William J. Clancey, Shane T. Mueller, Florian Jentsch, and Mohammadreza Jalaeian. 2023. ""Minimum Necessary Rigor" in empirically evaluating human–AI work systems." *AI Magazine* 1–8. https://doi.org/10.1002/aaai.12108

AUTHOR BIOGRAPHIES

Gary Klein received his Ph.D. in experimental psychology from the University of Pittsburgh, and is the Chief Scientist at Shadowbox LLC. He established the paradigm of Naturalistic Decision Making. His book Sources of Power was a New York Times best seller. He pioneered macrocognitive modeling and the creation of toolkits to help people make effective use of AI systems.

Robert R. Hoffman is an Emeritus Senior Research Scientist at the Florida Institute for Human and Machine Cognition. He received his Ph.D. in experimental psychology at the University of Cincinnati. He helped establish the principles of Human-Centered Computing and specializes in cognitive systems engineering for the design and analysis of macrocognitive work systems.

William J. Clancey is an Emeritus Senior Research Scientist at the Florida Institute for Human and Machine Cognition. He received his Ph.D. in Computer Science at Stanford University. He introduced the notion of heuristic classification and pioneered development of explanation and discourse systems using AI qualitative-modeling methods. His NASA Ames team developed voiced-commanded agent systems for Mars exploration, which received the Exceptional Software Award for supporting International Space Station operations.

Shane T. Mueller received his Ph.D. at the University of Michigan in Cognitive Psychology. He is a Professor of Psychology in the Department of Cognitive and Learning Sciences at Michigan Technological University. His research focuses on computational approaches to studying human memory, decision-making, and expertise, as well as human-centered approaches to AI.

Florian Jentsch obtained a Ph.D. in Human Factors Psychology and holds degrees in Aeronautical Sciences and in Aeronautics and Astronautics. He is a Professor of Psychology at the University of Central Florida. His research interests focus on training for high-consequence occupations, specifically pilots and the military, on team performance, and on human–systems interaction.

Mohammadreza Jalaeian received his Ph.D. in Learning Systems Design at the Southern Illinois University. He is currently a Research Engineer at the Ohio State University, Department of Integrated Systems, Cognitive Systems Engineering Lab. He specializes in human–machine teaming, complex macrocognitive systems, and visual analytics.