Human Factors in Energy: Oil, Gas, Nuclear and Electric Power, Vol. 54, 2022, 40–48
https://doi.org/10.54941/ahfe1002219

Operator Insights and Usability Evaluation of Machine Learning Assistance for Power Grid Contingency Analysis

John Wenskovitch[1], Alexander Anderson[1], Slaven Kincic[1], Corey Fallon[1], Danielle Ciesielski[1], Jessica Baweja[1], Molly C. Mersinger[2], and Brett Jefferson[1]

[1]Pacific Northwest National Laboratory, Richland, WA 99354, USA
[2]Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA

ABSTRACT

Introducing machine learning (ML) assistance into any established process comes with adoption barriers, including entrenched procedures, technological and human readiness levels, human-machine trust, and work culture resistance to change. These barriers are even greater in critical operations such as operating a national or regional power grid, in which both regulatory frameworks and the importance of maintaining reliability levels causes additional resistance to the adoption of new computational support. Developers of future systems and job aides must consider not only technical aspects, but also whether new systems are usable by power system operators. This work presents the methodology and results of a study to evaluate the usability and readiness of a prototype recommender system for power grid contingency analysis. We explore operator cognitive load and evaluate operator performance when solving a collection of scenarios both with and without recommender assistance. We also examine operator trust in the system. We report insights gained on the readiness of the system using a collection of evaluation techniques.

Keywords: Human-machine teaming, Power systems, Usability evaluation

INTRODUCTION AND BACKGROUND

Introducing automation into established workflows and processes is difficult. Factors that affect technology adoption can range from difficulties with novel systems (e.g., reliability and usability) to organizational constraints (e.g., access to training and data continuity) and more (Ertmer, 1999). Introducing machine learning (ML) presents still further challenges, including lack of trust in decision-making and results (Hoff and Bashir, 2015), loss of user autonomy (BenMessaoud et al. 2011), and increased workload during the adjustment process (Ludwick and Doucette, 2009).

One field that frequently struggles to incorporate such new technologies is power systems. Though the introduction of artificial intelligence into power system operations was first proposed in the 1980s

(Wollenberg and Sakaguchi, 1987), challenges with control room work culture tend to stifle technological innovation. However, recent thrusts into renewable energy may necessitate the introduction of advanced algorithms to better forecast generation in solar and wind energy (Makarov et al. 2011). Techniques such as reinforcement learning are generating interest for autonomous voltage control and smart grids (Zhang et al. 2018). Future power systems must consider not only technical aspects but also technology (Mankins, 1995) and human (American National Standards Institute, 2021) readiness levels (TRL and HRL), adherence to existing operational procedures, human-machine trust, and situational awareness of grid performance.

In this work, we present the methodology and results of a study to measure domain expert trust, elicit feedback, and understand technological usability and impact when a machine learning decision support assistant is introduced into contingency analysis for real-time power grid simulation. We evaluate the usability and readiness of a prototype recommendation system called ACAT (AI-Based Contingency Action Tool), a neural network which recommends contingency mitigations in order to quickly address potential power grid violations (Chen et al, 2019). Our study makes use of multiple data acquisition and processing techniques to evaluate the ACAT system, including assessing operator performance via analytical performance functions, inferring operator cognitive load with heart rate variability (HRV) data, and capturing operator feedback with both structured surveys and semi-structured debriefing interviews. We report the insights gained from these techniques separately and in combination, and we discuss their implications on the future of ML for power grid operations.

EXPERIMENT DESIGN

Our goal was to evaluate the current usability of ACAT. This system serves as a recommendation provider, displaying recommendations to remediate contingency analysis (CA) violations. An artificial neural network (ANN) and a semi-supervised corrective action algorithm display a sequence of control actions to the operator as a recommended procedure for resolving the contingency. This experiment was designed to replicate the naturalistic decision making processes used by power system operators first documented by Greitzer and Podmore (2008).

Control Room

The experiment was conducted in the Electricity Infrastructure Operations Center (EIOC) West Control Room at Pacific Northwest National Laboratory (PNNL). The EIOC provides a functional, configurable, and pragmatic simulated control room environment for transmission and distribution system operations and contains 16 operator consoles arranged at three control desks, connected to a dedicated network and server enclave. The control room also features a 12m x 3m multiplexing video wall system. The realistic control room environment provides opportunities for conducting studies

involving human factors, cognitive systems engineering, human-machine teaming, and evaluation of new systems and technologies relative to operator cognitive load.

Experiment Procedures

To simplify the experiment and focus on the role of the ACAT recommender, it was determined to use only a single participant per experiment who served in multiple control room roles (i.e., as reliability coordinator, balancing authority, and transmission operator). One of our primary goals was to provide an example of how early TRL technology can be evaluated to assess and improve the technology's HRL and TRL simultaneously. ACAT, at the time of this study, was at a TRL of 3. During each trial (which lasted roughly 10 minutes), the participant was responsible for identifying the contingency analysis violation, determining mitigating control actions, and then implementing those control actions on the real-time simulation of the power system. The experiment took place over the course of three partial-day sessions. The first day included a training session, so that the participant had time to work with the simulated grid and become familiar with the reliability of the ACAT software. Following the training session, the participant was presented with scenarios of varying difficulty and asked to resolve the contingencies in each scenario. These trials alternated between two experimental conditions: solving a scenario while having access to the ACAT recommender and solving a scenario without the ACAT recommender. Each of the scenarios was tested under both experimental conditions, but the participant did not see a scenario twice on the same day. Following each trial, the participant completed a survey appropriate for the trial condition.

Participant

Due to the significant time investment required to participate in the study, as well as COVID safety considerations, we provide an application of the methodology with a single representative user to assess the current HRL state of ACAT and identify improvement areas. Currently working as a Power System Research Engineer, our participant has a breadth of expertise in power systems, including real time control room operation, implementation, and deployment of real-time applications such as SCADA, State Estimator (SE), real time contingency analysis, real-time voltage and transient stability, cascading outages, and deployment of Energy Management Systems (EMS). The participant holds a PhD in power systems, has previously worked for multiple power utilities, and has experience in various operation and planning studies.

METHODOLOGY COMPONENTS

The methodology used in this study is intended to incorporate the rapid acquisition of objective (heart rate variability data and performance scoring computations), subjective (survey results), and qualitative (semi-structured interview) feedback from a participant. From this methodology, we are able to glean data used to measure domain expert trust, cognitive load, and the

technological usability and impact of a new system, as well as to identify potential correlations between these measures.

Heart Rate Variability

Participants are fitted with a Zephyr bioharness, a commercial device capable of measuring six inputs and reporting more than twenty biometrics, including heart rate variability (HRV). Because of the demonstrated connection between decreases in HRV and increases in cognitive load (Aasman et al. 1987), HRV is used as a cognitive load proxy throughout the duration of an experimental trial. Data is collected from the bioharness using the accompanying OmniSense software. The OmniSense software corrects for outliers and calibrates to the wearer. Electrocardiogram (ECG) data is averaged for comparison across the two experimental conditions.

Performance Scoring

In this domain, there can be many ways to solve a given contingency. The dimensions in which *performance* is measured include time to solve, severity of remaining violations, financial costs to take mitigating actions, and downstream consequences that may include contributing to a future grid issue. Since our aim with this human factors work overall is to provide a generalizable framework, we measure factors that are not particularly tied to the power systems industry and that have some standing in the human factors literature as generally meaningful to performance. These metrics include completion times, number of actions taken per evaluation, and number of evaluations per trial. These behavioral measures have ties to cognitive load measures, workload measures, and often are used to determine task difficulty. Here, we primarily use them to support findings from other measures.

Surveys

Participants receive a modified version of Madsen and Gregor's Trust Questionnaire (Madsen and Gregor, 2000), with versions of the questionnaire tailored for the two experimental conditions. This self-report questionnaire measures several constructs believed to underlie trust in a system. According to Madsen and Gregor, a system is trusted if participants perceive it as understandable, technically competent, and reliable, comprising the Cognitive-Based component of trust. In addition, participants who have faith in a system and a personal attachment to the system will be more likely to trust the technology, comprising the Emotion-Based component of trust. For each item, participants are asked to rate their agreement with a statement on a 5-point Likert scale. Additionally, several items are included in the questionnaire to assess the importance of each construct in participant trust.

Post-Experiment Interview

Throughout the experiment, we collect qualitative feedback from participants as they comment on their actions and describe their approach towards solving a scenario. Because we are measuring cognitive load, participants are not explicitly asked to follow any think-aloud procedures, but we interact with them as they work through each study trial. Following the study, we

also conduct a short debrief, gathering additional feedback on the technological readiness of the system. This information is aggregated and summarized to provide structured feedback to developers working on the next iteration of the system.

RESULTS AND DISCUSSION

The sections below describe the results of each component of the methodology. Notably, we find common themes that connect the various components of our methodology. For example, the slight increase in cognitive load identified in the HRV data is also explained by distrust in the ACAT recommendations identified in the survey results, while the qualitative feedback indicating that the participant often disregarded the ACAT recommendations is supported by a lack of performance differences between the two experimental conditions.

Heart Rate Variability (HRV)

This HRV analysis seeks to identify periods of increased cognitive load through observation of lower variation in heart rates. The observed trials began with average heart rate standard deviations ranging from 35-50 milliseconds, which typically stayed in that range for the first 2-3 minutes of the trial. Both the with and without ACAT data sets show clear decreases in variability (and hence increases in cognitive load) as tasks went from the three-minute to the five-minute mark (see Figure 1).

Overall, the HRV scores showed a slight increase in cognitive load earlier while using the ACAT recommender. Ideally, introducing a recommender into a workflow would decrease cognitive load rather than increase it. As seen in the qualitative feedback, this could be due to a lack of operator comfort with the tool, distrusting its recommendations and thinking too critically about their quality and effectiveness. Operator training on future versions of ACAT could change this result. Additionally, using the same participants in these studies across multiple development and evaluation cycles can benefit both the participant and the study, as the participant will come into the study with some previous knowledge of system capabilities and prior quality.

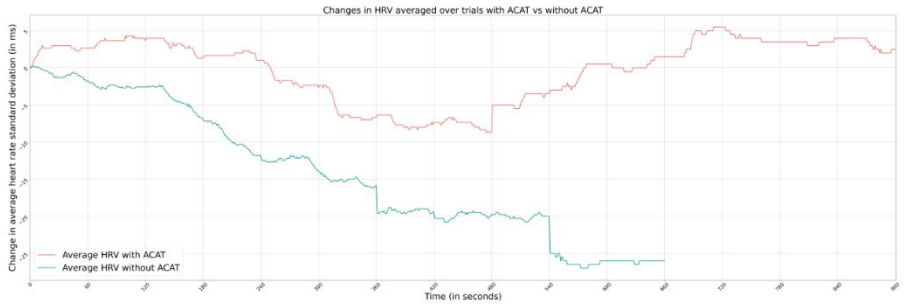

Figure 1: Participant heart rate variability averaged over all trials, separated into the with- and without-ACAT experimental conditions.

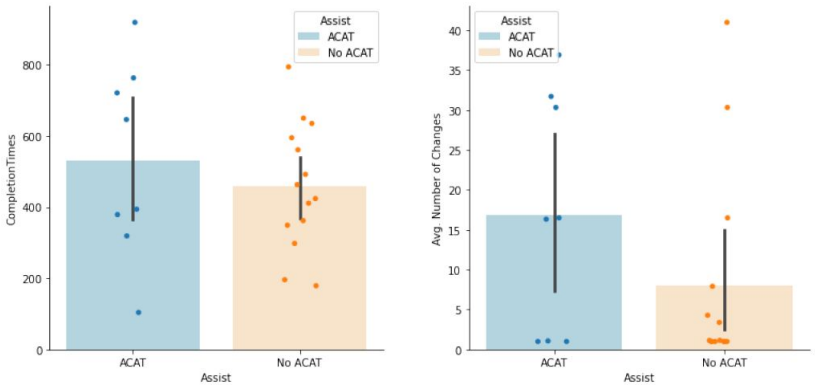

Figure 2: Completion times and average numbers of actions taken by the study participant in both experimental conditions.

For tasks completed in less than five minutes, the HRV generally maintained a standard deviation of at least 30 milliseconds. This indicates that quick completion of tasks staved off a heavy cognitive load. If operators can trust ACAT sufficiently to accept and implement its recommendations quickly, overall cognitive load for operators may decrease across their full workflow. After the five-minute mark, the cognitive load remained heavy on average, corresponding with the implementation phase during which an operator must communicate the mitigation steps to others. Once a solution had been reached, the workload was similar with and without ACAT, which is unsurprising since ACAT no longer plays a role in the workflow at this point.

Performance Scoring

We collected completion times, number of actions per evaluation, and total number of evaluations per scenario/trial. While on average the participant was slightly faster without ACAT, this difference is not statistically significant. Similarly, there was no noticeable difference between ACAT scenarios and no ACAT scenarios with respect to number of evaluations. When average number of actions per evaluation is considered, we again find a slightly fewer number of average actions for no ACAT scenarios, but not enough to drive a statistical difference.

During the two days of testing with our participant, there is behavioral evidence that suggests he preferred to resolve contingencies using the operations manual and his tacit knowledge rather than relying on the ACAT recommender. The scoring analysis indicates little difference between participant performance in both experimental conditions (see Figure 2). The incorporation of ACAT into the contingency analysis workflow did not drastically improve or degrade the operator's performance. This lack of statistical differences is partially due to the participant's tendency to disregard the ACAT recommendations, which is evidenced by the trust results from the survey and the qualitative feedback.

The participant indicated that short completion times in later trials were due to his progressively greater familiarity with the simulated grid over the

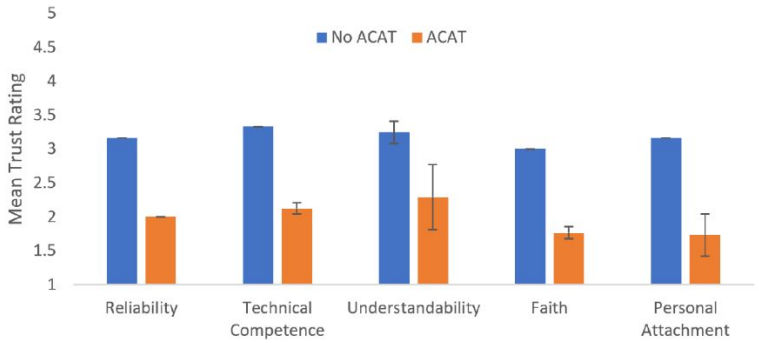

Figure 3: Mean rating across all trials for each cognitive- and emotion-based construct of trust as a function of experimental condition (with or without ACAT recommendations).

course of the experiment, not an effect of seeing the ACAT recommendation when this scenario was presented on the previous day. It is likely that additional training and exposure to this grid by experienced participants could lead to additional efficiency boosts in other scenarios during future experiments.

Survey Outcomes

Trust results from the survey indicated that our participant had an overall lower trust in the ACAT recommendations than in his personal expertise and in the operational procedures. This also matches the qualitative feedback that was captured; the participant did not feel that he had sufficient exposure to ACAT to judge its reliability. In five of the six scenarios, trust in ACAT was lower than trust in Operations Procedures for all constructs (see Figure 3).

Only two scenarios did not reveal lower trust in ACAT than in Operations Procedures across trust constructs. In one case, this was due to low trust in the procedures for resolving the contingency. In the other, Emotion-Based trust remained lower for ACAT (particularly seen in the Faith construct) whereas Cognitive-Based trust was comparable. An explanation for lower Emotion-Based trust ratings in this scenario may be the need for more exposure. The time needed to develop positive affect toward a system may be greater than the time needed to perceive that a system is functioning reliably, accurately, and is understood.

The importance of each trust construct was compared to its mean rating for ACAT. The results reveal that Personal Attachment is less important than other constructs in contributing to overall trust in ACAT. In all cases, the mean construct rating (ranging from 1.7 to 2.3) was lower than the perceived importance of the construct to overall trust (ranging from 3.0 to 4.0). The construct Understandability varied the most across items for all six scenarios. Overall, the survey results suggest it may not be enough to simply improve system transparency as a technique for building trust. Developers should also consider improving performance in order to increase trust.

Qualitative Feedback

The semi-structured debrief captured participant feedback about both the current state of ACAT and the struggles that the participant encountered

when considering the system recommendations during the experiment trials. With respect to the tool, the participant identified several usability issues that were not considered by the developers. Most notably, ACAT used bus numbers rather than substation names. Because operators are more familiar with the names, they would encounter additional work in converting the ACAT recommendation into an actionable recommendation. Recommendations generated by ACAT also differed from common operational practices, leading to lower trust in the system recommendations. This tendency to trust the operational procedures and personal experience over the recommendations matches the trust measures uncovered in the survey results.

DISCUSSION

The discussion around artificial intelligence in the workplace often centers on the capabilities of the machine, proposed benefits to user workflows, and assumes increased efficiency. Our work aims to balance the conversation by emphasizing the human teammate/user and the very real, but often not thought about challenges to adopting and using semi-autonomous tools. This work provides an example of how to measure these challenges and identify gaps in tool development as they relate to the end-users that will ultimately be responsible for deploying them. The authors realize that for society, the introduction of technology has often meant replacing human workers with autonomous ones and that livelihoods are impacted by the tremendous push for more automated procedures. We hope that our work shows that humans are a critical component to the adoption of many proposed AI technologies.

Limitations and Future Work

In this study, we evaluated the ACAT system using a single participant. The availability of control room operators has been extremely limited due to COVID and the critical nature of Power Grid operations. Our research team does plan to conduct a more robust study with several engineers using the methodology. Our performance metrics were also coarse and aimed only to provide a high-level view of our participant's mitigation of power system violations. The authors note that more nuanced metrics can be developed for this specific application domain. The ACAT tool used for this study revealed serious limitations in its ability to consider other aspects of decision making for violation mitigation. ACAT was biased toward actions that did not match operational procedures, and the system usability would be improved with closer connections between developers and operators.

CONCLUSION

This work presents a methodological approach applied to evaluating an AI-Based Contingency Action Tool (ACAT), a recommendation system designed to aid power system operations in their decision-making process for resolving potential violations in power grid equipment and state. We capture objective, subjective, and qualitative measures during a controlled experiment with two conditions: the operator's standard workflow that makes use of personal

expertise and an operations manual vs. the inclusion of the AI-based recommender to provide mitigation options. As a result of our study, we identified several key issues with the ACAT system, ranging from usability concerns to mismatches between the system recommendations and current operational procedures. Collecting this feedback rapidly and delivering it to developers can aid in accelerating the development and evaluation cycle of ACAT, assisting in its deployment into control rooms sooner than the typical slow pace of novel applications in this domain space.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. Yousu Chen, Ms. Lyndsey Franklin, and Mr. Blaine Mcgary for their contributions to this project.

REFERENCES

Aasman, J., Mulder, G. and Mulder, L.J., 1987. Operator Effort and the Measurement of Heart-Rate Variability. *Human Factors*, *29*(2), pp. 161–170.

American National Standards Institute (ANSI) and Human Factors and Ergonomics Society (HFES), 2021. Human Readiness Level Scale in the System Development Process (ANSI/HFES 400-2021).

BenMessaoud, C., Kharrazi, H. and MacDorman, K.F., 2011. Facilitators and Barriers to Adopting Robotic-Assisted Surgery: Contextualizing the Unified Theory of Acceptance and Use of Technology. *PloS One*, *6*(1), p. e16395.

Chen, Y., Yin, T., Huang, R., Fan, X. and Huang, Q., 2019, December. Big Data Analytic for Cascading Failure Analysis. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 1625–1630). IEEE.

Ertmer, P.A., 1999. Addressing First- and Second-Order Barriers to Change: Strategies for technology integration. *Educational Technology Research and Development*, *47*(4), pp. 47–61.

Greitzer, F.L. and Podmore, R., 2008. *Naturalistic Decision Making in Power Grid Operations: Implications for Dispatcher Training and Usability Testing* (No. PNNL-18040). Pacific Northwest National Lab.(PNNL), Richland, WA.

Hoff, K.A. and Bashir, M., 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), pp. 407–434.

Ludwick, D.A. and Doucette, J., 2009. Adopting Electronic Medical Records in Primary Care: Lessons Learned from Health Information Systems Implementation Experience in Seven Countries. *International Journal of Medical Informatics*, *78*(1), pp. 22–31.

Madsen, M. and Gregor, S., 2000, December. Measuring Human-Computer Trust. In *11th Australasian Conference on Information Systems* (Vol. 53, pp. 6–8). Brisbane, Australia: Australasian Association for Information Systems.

Makarov, Y.V., Etingov, P.V., Ma, J., Huang, Z. and Subbarao, K., 2011. Incorporating Uncertainty of Wind Power Generation Forecast into Power System Operation, Dispatch, and Unit Commitment Procedures. *IEEE Transactions on Sustainable Energy*, *2*(4), pp. 433–442.

Mankins, J.C., 1995. Technology readiness levels. *White Paper, April*, *6*(1995).

Wollenberg, B.F. and Sakaguchi, T., 1987. Artificial Intelligence in Power System Operations. *Proceedings of the IEEE*, *75*(12), pp. 1678–1685.

Zhang, D., Han, X. and Deng, C., 2018. Review on the Research and Practice of Deep Learning and Reinforcement Learning in Smart Grids. *CSEE Journal of Power and Energy Systems*, *4*(3), pp. 362–370.