The need for a socio-technical systems testbed to support characterization and evaluation of AIES behavior

Systems Engineering Research Center (SERC) / Acquisition Innovation Research Center (AIRC)

Submitters:
Zoe Szajnfarber (SERC Chief Scientist/GWU) and Dinesh Verma (SERC/AIRC Executive Director/Stevens)

Contributors (SERC/AIRC):
Peter Beling (Virginia Tech), David Broniatowski (GWU), Dan DeLaurentis (Purdue), Laura Freeman (Virginia Tech), Erica Gralla (GWU), Jitesh Panchal (Purdue)

Thank you for the opportunity to provide input through this Request for Information (RFI) Related to NIST's Assignments under Sections 4.1, 4.5 and 11 of the Executive order Concerning Artificial Intelligence (Sections 4.1, 4.5, and 11). We are specifically responding to item #2 sub bullet 4: "Applicability of testing paradigms for AI system functionality, effectiveness, safety, and trustworthiness including security, and transparency, including paradigms for comparing AI systems against each other, baseline system performance, and existing practice, such as: Model benchmarking and testing, and Structured mechanisms for gathering human feedback, including randomized controlled human-subject trials, field testing, A/B testing, AI red-teaming." Our input is particularly relevant to NIST's efforts related to the Assessing Risks and Impacts of AI (ARIA) environment.

Understanding AI System behavior requires that the *system* be evaluated in a *representative* environment, which increasingly includes the interactions of many technical components, some of them AI enabled, and many actors, be they individuals, organizations, or institutions. Existing practices for AI model verification typically focus on the model as the system and aim to stress test it across anticipated operating conditions; this approach is fundamentally limited in its ability to capture emergent behaviors that arise a) because of the stochastic nature of AI algorithms and b) through the complex interactions among and between multiple actors and systems, over time. To overcome these limitations, we believe that there is a need to conceive of and build a "socio-technical systems test bed" that can serve as a national resource to a) develop general safety guidelines for classes of AI Enabled Systems (AIES) of Systems and 2) serve as a "test range" for federal agencies wishing to evaluate a system before acquisition or deployment and/or train operators on these new types of systems.

The notion of *representativeness* is core to the feasibility of a socio-technical systems testbed. For a testbed to usefully inform policy, regulation, and practice, it must be able to replicate key behaviors of the full-scale deployed system, but to do so on a smaller scale and in a way that is readily instrument-able – that is, in which relevant data can be collected. Many of the safety and performance questions we wish to evaluate manifest at very large scales – through interactions with society itself or at least with large groups and expensive and/or deployed systems; this

makes it impractical, if not infeasible, to experiment on the full system. The question is: how much of the real-world system needs to be replicated to achieve ecological validity? Depending on the evaluation intent, it may be possible to drastically reduce the complexity of the system under study, but it is critical to understand what aspect of the system of systems need to be represented to support the desired inference.

Other disciplines already do this. For example, in earthquake engineering, the vibration loads experienced by a building are highly non-linear and it is understood that important dynamics can't be adequately studied, even in high-fidelity simulations, without some amount of hardware in the loop. That field has devised a type of "hybrid simulation" that typically includes a realistic building joint that receives the vibration load but captures the rest of the building as a simulation that provides physical feedback on that joint. Relatedly, in biology, researchers have established guidelines for which types of questions can be asked by experimenting with a cell-line vs. a rat "model" vs. clinical trials. In both cases, the key is matching the question to an appropriate "model world" which enables researchers to balance costs with the need to observe certain kinds of dynamics in more realistic settings.

There would be enormous benefit to the community if such standards and best practices existed for emerging classes of AI and AIES. To make progress towards this goal we suggest a need for three types of activities:

1) Formation of a community of practice. Multiple federal agencies, corporations and academics are currently wrestling with similar needs. While there are certainly unique aspects of each problem domain, many of the system dynamics are similar enough that shared testbeds may be able to support multiple needs. Further, there are opportunities to learn across disciplinary practices, for example, comparing war gaming and tabletop simulations, to classes of hardware-in-the-loop simulations.
2) Fundamental research characterizing the features of a testbed needed for representativeness. At a minimum, this testbed must be able to capture the behaviors resulting from socio-technical interactions with AIES, but what aspects of scale matter? For example, how large does a network of users need to be to replicate the dynamics of the billions of users on social media platforms, given the importance of subnetwork interactions? How much of the structure of their environment needs to be replicated for observed behaviors to represent those of the real system? There is potential to discover scaling properties that can inform test bed design in general.
3) Prototype testbeds to support proof-of-concepts: Test beds are context dependent, and it is likely necessary to pick several initial systems that vary in key features. Systems that emphasize information flows (e.g., on social media platforms) likely benefit from different abstractions, and associated testbed features, than those that emphasize physical flows (e.g., autonomous drone swarms) from those emphasizing operations (e.g., delivering physical goods). Working on prototypes will shed important light on aspects of design, but also equally important, strategies for managing these types of test beds for use by multiple actors, in ways that support learning across efforts.

The Systems Engineering Research Center (SERC), with the Acquisition Innovation Research Center (AIRC) added more recently, are particularly well positioned to support all these efforts. The SERC is the only University Affiliated Research Center (UARC) hosted directly by the Office of the Under-Secretary of Defense (Research and Engineering), and represents a network of 25 universities and all the minority serving institutions. Over the course of more than a decade, faculty and researchers from the various SERC/AIRC universities have developed strong connections across multiple federal government agencies and industry, along with nucleating a strong and collaborative network of university researchers. This has enabled significant transition of academically rigorous research to practice. The UARC contractual mechanisms make it possible to coordinate effectively across stakeholders in a relatively easy manner (A UARC is a sole source IDIQ contractual mechanism for all research that falls within its charter). In the last academic fiscal year alone, SERC and AIRC had more than $35M in research expenditures, leveraging effort by more than 200 faculty/staff members across the research network. SERC/AIRC has a long track-record of supporting each of the types of activities described above, including specific expertise in the context of AIES. As an example of each:

1. Formation of community of practice: AIRC is currently building a government-industry-academia community around Digital Material Management. This is the Department of the Air Force's strategy to leverage the Digital Transformation across all the classical acquisition functions – contracting, program management, engineering, test and evaluation, logistics and material readiness, and so on. AIRC was asked to facilitate the substantive interaction of diverse stakeholders with a goal of converging on a roadmap and shared vision for how to move forward. As evidence of SERC/AIRC's convening power, the first workshop filled the 200 available seats almost immediately and included senior representatives from major stakeholders across government, industry and academia. Topically more closely related, over the last four years, SERC has built a community around Systems Engineering for AI and AI for Systems Engineering. Starting with a small workshop in collaboration with the US Army DEVCOM Armaments Center, the annual SE4AI/AI4SE workshop is now a 200 person annual event and a key outlet for related work.
    - Sample sharing website: DMM Industry Association Consortium: https://guide.dafdto.com/industry-association-consortium-sharing-site/
    - Sample report: AI4SE & SE4AI Research and Application Workshop Summary Report, Sept 27-28, 2023, https://sercuarc.org/event/ai4se-se4ai-workshop-2023/
2. Fundamental Research on Representativeness of Testbeds: SERC/AIRC researchers have led past projects examining foundational issues related to representativeness of experimental setups (testbeds) for research in related domains. For example, a team from GWU-Purdue-Stevens conducted a study on the representativeness of model worlds for design research. Projects often aim to make inferences on how new design tools will affect the way engineering organizations innovate, which is a phenomenon that is highly contextual to organizational culture and lives in a large network of thousand-person engineering teams collaborating over extended periods of time. The team proposed a framework of subject-context-task interactions that could be simplified together without losing representativeness of system behaviors. For example, rather than focusing on the

validity of studying classroom design tasks, they shifted the focus to calibrating the task to the context and experience of the subjects. They also explored human-simulation hybrids to capture different aspects of the dynamics.

- Sample publication: Szajnfarber, Zoe, et al. "A call for consensus on the use of representative model worlds in systems engineering and design." *Systems Engineering* 23.4 (2020): 436-442.

3. Testbed proofs of concept: SERC/AIRC researchers have been developing a "test harness" for the Department of Defense that provides data, AI models, and Test and Evaluation (T&E) capabilities in a computational environment that support empirical study of data strategies and plans for test. Led by Virginia Tech, a core aspect of the strategy is that it aims not only to accelerate and improve T&E activities in the short term, but also support workforce training and cross-pollination through its data repository and training modules. While this project is focused on the model pipeline, other efforts are looking at extending this to a cyber-physical system with users in the loop. Model-based systems engineering, physical, and other approaches will be used to model the subject system from a variety of perspectives (e.g., control, human-system integration, communication) and at varying fidelities. The collection of models will support the study of the importance of system understanding in T&E of AIES.
    - Sample publication: Freeman, Laura J., et al. "Digital Engineering Enhanced T&E of Learning-Based Systems." Acquisition Research Program Conference, 2023.