**Title**: Evaluation and Design of Generalist Systems (EDGeS)

**Authors**: John Beverley ([johnbeve@buffalo.edu](mailto:johnbeve@buffalo.edu))

The field of AI has undergone a series of transformations, each marking a new phase in its development. The initial phase emphasized curation of symbolic models which excelled in capturing reasoning but were fragile and not scalable. The next phase was characterized by machine learning models – most recently large language models (LLMs) - which were more robust and easier to scale but struggled with reasoning challenges. Now, we are witnessing a return to symbolic models as complementing traditional machine learning. Successes of LLMs are contrasted with their inscrutability, inaccuracy, and hallucinations, which underwrite concerns over the reliability and trustworthiness of these systems, motivating investigations into commonsense reasoning, AI explainability, and formal verification techniques. Proper assessments of hybrid machine learning/symbolic systems require novel strategies to facilitate comparisons of performance and guide future AI progress. The EDGeS AAAI 2023 Spring symposium brought together researchers focusing on novel assessments and benchmarks for evaluating hybrid and artificial general intelligence (AGI) systems. This symposium revealed what was already suspected: research concerning evaluation of machine learning/symbolic hybrids and AGI is, unfortunately, lacking. Even so, the discussion was fruitful.

One major theme that emerged was the exploration of symbolic reasoning systems and LLMs as complementary technologies. Grant Passmore (Imandra) illustrated how GPT-4 could be employed to generate symbolic representations of financial policy documents in natural language, which could be ingested by a proof assistant and model checker to identify potential loopholes in financial algorithms. Michael Gruninger (University of Toronto) outlined at how we might characterize presumed sets of rules governing inputs/outputs of LLMs, making them easier to understand and validate. Ramesh Bharadwaj (Naval Research Laboratory) encouraged coupling LLMs with symbolic AI in the interest of automating code vulnerability detection. Optimism was met with challenges on at least two fronts. First, the adequacy of symbolic representations differs by domain. Financial policies may be amenable to rigorous formal representations, but formally representing even mundane activities, such as cracking an egg, can be notoriously laborious and often left incomplete. Second, many find it challenging to trust the outputs of model checkers and proof assistants for anything substantial. So, while exploration into relationships between LLMs and symbolic reasoning systems is growing, concerns of generalizability and trust are sobering realities.

A further theme centered on how consciousness, intelligence, and commonsense reasoning relate to AGI. Leora Morgenstern (PARC) reported on the defeat of the Winograd Schema Challenge by LLMs. This challenge was designed around pairs of sentences involving pronoun reference ambiguity, which appear to require commonsense reasoning to disambiguate. The success of LLMs at this task led Leora to question the role of surrogate task AGI testing. Joscha Bach (Intel) presented a framework of capabilities relevant to AGI, characterized by reflective awareness of input learned from embodied perception alongside internal validation by reasoning, and creative, autonomous interaction with the environment. Similarly, Pei Wang of Temple University, re-imagined intelligence as the ability to adapt to one's environment while operating with insufficient knowledge and resources. Gadi Singer (Intel) later added that one milestone we must reach before obtaining AGI is the development of 'cognitive AI', consisting of systems with multi-modal reasoning, learning, and unlearning capabilities. Joscha, Pei, and Gadi each highlighted important characteristics differentiating AGI from task-focused AI systems. Altogether, these discussions made clear that AGI poses unique challenges for evaluation and assessment that go beyond measuring performance of a specific task. Rather, assessing AGI systems will be more like assessing natural, embodied intelligence.

  Though the main theme of our symposium, there were few concrete discussions of novel strategies for evaluating and benchmarking either machine learning/symbolic hybrids or AGI. Even so, speakers throughout the symposium reiterated the importance and lack of existing assessments and benchmarks for each. Indeed, many felt such research is a necessity, as we explore this new phase of AI development.

Joscha Bach, Amanda Hicks, and the late John Piorkowski co-chaired the event, with Tetiana Grinberg, John Beverley, Steven Rogers, Grant Passmore, Ramin Hasani, Casey Richardson, Richard Granger, Jascha Achterberg, Kristinn Thorisson, Luc Steels, and Yulia Sandamirskaya as co-organizers. Papers from the symposium are published in *OpenReview*. The report was written by John Beverley (Assistant Professor, University of Buffalo).