International Journal of Artificial Intelligence (IJAI)
Volume 4, Issue 2, July - December, 2023, pg 1-5.
https://ijai.in/

Defining Metrics for Evaluating General Artificial Intelligence Across Diverse Problem Domains

Luv M Eldho,
USA.

Citiation: Luv M Eldho. (2023). Defining Metrics for Evaluating General Artificial Intelligence Across Diverse Problem Domains. *International Journal of Artificial Intelligence*, *4*(2), 1–5.

ABSTRACT

Defining robust and universal metrics for evaluating General Artificial Intelligence (GAI) is essential for its development and implementation across diverse problem domains. This paper explores the theoretical and practical aspects of GAI evaluation, focusing on frameworks that assess adaptability, problem-solving capability, and generalization. Existing benchmarks often emphasize narrow tasks, failing to capture the broader spectrum of intelligence characteristics. By integrating insights from psychology, neuroscience, and computational theory, this work proposes a multidimensional evaluation model incorporating metrics such as knowledge transferability, reasoning depth, and robustness against novel challenges. The study also discusses the importance of domain-agnostic evaluation standards to ensure fairness and comprehensiveness.

KEYWORD

General Artificial Intelligence, Evaluation Metrics, Adaptability, Problem-Solving, Generalization, Benchmarking, Domain-Agnostic, Intelligence Testing

1.Introduction:

The development of General Artificial Intelligence (GAI) presents one of the most profound and ambitious goals in the field of computer science and cognitive modeling. Unlike narrow AI, which focuses on domain-specific tasks, GAI aspires to replicate the adaptive, context-sensitive, and transferable intelligence displayed by humans across a broad array of domains. Evaluating such a complex system requires metrics that go beyond traditional performance benchmarks. The challenge lies in formulating criteria that can capture the essence of general intelligence — including learning efficiency, generalization across tasks, and resilience in unfamiliar contexts.

While many existing benchmarks emphasize narrow capabilities, such as image classification or game performance, they fail to capture the core attributes of general intelligence. These shortcomings underscore the urgent need for domain-agnostic and

comprehensive evaluation standards. This paper explores theoretical foundations and proposes a multidimensional model to address this gap. By drawing on insights from psychology, neuroscience, and algorithmic information theory, we aim to redefine how GAI should be assessed in a robust and universally applicable manner.

2. Literature Review

The concept of evaluating machine intelligence is not new. Alan Turing's seminal work introduced the idea of an "Imitation Game," later known as the Turing Test, to assess a machine's capability to exhibit behavior indistinguishable from that of a human (Turing, 1950). Although influential, the Turing Test has been criticized for focusing on surface-level mimicry rather than true cognitive flexibility. Later thinkers such as McCarthy and Hayes (1969) and Newell and Simon (1976) emphasized the role of symbolic reasoning and search mechanisms in intelligent behavior, laying groundwork for more structured approaches.

Building on this foundation, scholars such as Marcus Hutter (2005) and Shane Legg (Legg & Hutter, 2007) proposed the idea of "universal intelligence," rooted in algorithmic information theory. This formulation considers an agent's ability to succeed across a broad set of environments, weighted by their simplicity. While theoretically elegant, these models are difficult to compute in practice, pointing to the need for more operationally feasible metrics that maintain theoretical rigor without becoming computationally intractable.

3. Current Benchmark Limitations and the Case for Domain-Agnostic Metrics

Current evaluation frameworks often fall short because they are designed with narrow tasks in mind. Benchmarks like ImageNet, the Atari Learning Environment, or even more sophisticated platforms like OpenAI’s Gym and DeepMind’s Control Suite assess performance in limited, structured environments. These domains, while useful for measuring specific abilities like pattern recognition or motor control, fail to test broader capacities like knowledge abstraction, analogical reasoning, or cross-domain transfer learning.

Moreover, most benchmarks are static and lack adaptability. They do not evolve to test an agent’s capacity to respond to novel challenges — a critical feature of general intelligence. This rigidity incentivizes overfitting, where systems are engineered to excel in specific tasks without demonstrating genuine adaptability. Therefore, a new class of benchmarks must prioritize **domain-agnostic** evaluation to encourage systems that are not just competent in narrow settings but capable of operating across a diversity of environments and problems.

4. Proposed Multidimensional Evaluation Framework

To address these limitations, we propose a **multidimensional evaluation model** that emphasizes key characteristics of general intelligence. First among these is **knowledge transferability** — the ability of a system to apply learned information from one domain to another. This metric evaluates whether an AI system can abstract principles and use them in unfamiliar contexts, similar to how humans generalize prior learning.

Another critical metric is **reasoning depth**, which measures the agent's ability to perform multi-step inference and resolve complex, ambiguous scenarios. Systems that exhibit deeper reasoning are better equipped to handle real-world situations, where information is often incomplete or contradictory. Lastly, **robustness to novelty** — the AI's capacity to remain functional and accurate in the face of previously unseen data or environments — serves as a litmus test for generality. Together, these metrics provide a comprehensive lens for evaluating GAI beyond the constraints of domain-specific benchmarks.

5. Interdisciplinary Insights and Integration

To develop these metrics meaningfully, the field must draw on a variety of disciplines. Psychology, for instance, offers models of cognitive development and intelligence testing — such as Piaget's stages of development or modern psychometric tools — which can inspire analogous evaluations for machines. Neuroscience provides insights into brain adaptability and distributed processing, offering a biological parallel to modularity and plasticity in AI systems.

From the computational theory perspective, algorithmic complexity and information theory offer tools to formalize concepts like simplicity, compression, and learning efficiency. Integrating these insights allows for more nuanced and theoretically grounded assessment tools. For example, evaluating an AI's ability to compress and generalize across data types can serve as a proxy for its reasoning capabilities. This interdisciplinary approach ensures that evaluation metrics are not only practical but also grounded in scientifically validated understandings of intelligence.

6. Conclusion and Future Directions

Evaluating General Artificial Intelligence necessitates a shift from narrow benchmarks to more holistic and flexible metrics. Our proposed model emphasizes adaptability, reasoning depth, knowledge transfer, and robustness — dimensions that capture the essence of general intelligence across problem domains. Such a shift

would help ensure that the development of GAI aligns more closely with the ultimate goal of achieving machines that can understand, reason, and adapt like humans.

In the future, creating standardized and open-source benchmark platforms that dynamically evolve based on agent behavior could further enhance evaluation efforts. These platforms should simulate real-world complexity and ambiguity, encouraging the development of genuinely intelligent systems. Ultimately, the creation and adoption of robust, domain-agnostic evaluation standards will be a cornerstone in the responsible and effective advancement of General Artificial Intelligence.

References

1. Turing, Alan M. "Computing Machinery and Intelligence." *Mind*, vol. 59, no. 236, 1950, pp. 433–460.
2. McCarthy, John, and Patrick J. Hayes. "Some Philosophical Problems from the Standpoint of Artificial Intelligence." *Machine Intelligence*, vol. 4, 1969, pp. 463–502.
3. Newell, Allen, and Herbert A. Simon. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the ACM*, vol. 19, no. 3, 1976, pp. 113–126.
4. Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.
5. Hutter, Marcus. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer Science & Business Media, 2005.
6. Legg, Shane, and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines*, vol. 17, no. 4, 2007, pp. 391–444.
7. Schank, Roger C., and Robert P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, 1977.
8. Wang, Pei. *Rigid Flexibility: The Logic of Intelligence*. Springer Science & Business Media, 2006.

9. Yudkowsky, Eliezer. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, Oxford University Press, 2008, pp. 308–345.
10. Lake, Brenden M., et al. "Building Machines that Learn and Think Like People." *Behavioral and Brain Sciences*, vol. 40, 2017, e253.