Bitter Lesson of the ARC-AGI Challenge: Intelligence may look very different in machines and humans

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Abstraction and Reasoning Corpus (ARC) and the associated ARC-AGI challenge are benchmarks for evaluating core reasoning skills. Recently, OpenAI's new model *o*3 solved ARC-AGI, reigniting debate on whether such achievements reflect true reasoning or mere pattern matching. Rich Sutton's "Bitter Lesson" suggests that general methods at scale outperform specialized ones; in ARC, top solutions heavily rely on data augmentation. We examine this tension and propose that new concepts or metaphors may be needed to describe machine reasoning. This position paper argues that reasoning and intelligence in machines can differ fundamentally from human cognition: recognizing that intelligence is task-specific and diverse across animals, humans, and machines could lead to more appropriate benchmarks.

1 Introduction

The Abstraction and Reasoning Corpus (ARC) and the associated ARC-AGI challenge Chollet [2019] are benchmarks for core reasoning and abstraction in AI. These tasks require minimal prior knowledge and emphasize generalisation. Recently, OpenAI's new model *o*3 solved many ARC-AGI puzzles, reigniting debate on whether such achievements reflect genuine reasoning or mere pattern-matching. Rich Sutton's "Bitter Lesson" Sutton [2019] suggests that general methods at scale outperform specialized ones; indeed, many ARC solutions rely on data augmentation and test-time heuristics.

This position paper argues that reasoning and intelligence in machines may differ fundamentally from those in humans. Recognizing intelligence as task-specific and diverse among animals, humans, and machines may lead to more appropriate benchmarks.

2 Defining Artificial General Intelligence (AGI)

ARC emphasises adaptability to novel problems, aligning with the definition of intelligence as efficiency of skill acquisition across tasks Chollet [2019]. ARC tasks present input-output puzzles on small colored grids (each cell one of ten colors); the goal is to produce a correct output grid from limited examples, demonstrating generalisation. These puzzles rely on basic core priors (object permanence, goal-directedness, counting, basic geometry) that humans naturally possess. ARC avoids reliance on acquired cultural knowledge (e.g., language) to ensure a fair comparison between human and machine intelligence.

OpenAI's *o*3 model recently achieved around 87% accuracy on the ARC test, exceeding the 85% threshold often associated with artificial general intelligence (AGI). However, much of the evaluation set had previously been accessible through APIs, raising concerns about data contamination. As a result, it is still uncertain whether ARC-AGI has been solved sincerely.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Goodhart's Law states that once a measure becomes a target, it ceases to be a reliable measure. If ARC is optimized through data augmentation or heuristics, its ability to assess general reasoning may diminish. To address this, we propose testing robustness by introducing task variants: for example, modifying input-output examples or creating concept variants (e.g., ConceptARC tasks Mitchell et al. [2023]) to ensure methods capture underlying principles rather than shortcuts.

3 Machines, reasoning, and the need for neologisms

A key question is what constitutes genuine reasoning in machines. Are we simply projecting human metaphors onto machine behaviour? Terms like "reasoning", "understanding", and "abstraction" are laden with human-centric assumptions. Machines may operate by entirely different principles. To describe these emergent behaviours, new terms (neologisms) may be needed: for example, we could reconceptualize "artificial reasoning" as "mechanical abstraction" or "synthetic inference".

4 Generalisability to other datasets and other domains

AI systems that perform well on ARC tasks may still rely on task-specific shortcuts. To truly evaluate generalisation, such models should be tested on related tasks in different domains. For example, a system that solves an object-centric ARC puzzle could be tested on a maze navigation task requiring the same objectness reasoning (e.g., the PUZZLES benchmark Estermann et al. [2024]). Likewise, solving an ARC permutation task should translate to solving word-based permutation puzzles. By evaluating models on such analogous tasks, we can assess whether the underlying reasoning principles generalise.

5 A call for broader agreement on intelligence

"Intelligence" remains a "suitcase word" with many conflicting definitions. AI researchers often use functional definitions (e.g., goal achievement or optimisation), while cognitive scientists and neuroscientists may emphasise creativity, emotion, or embodied cognition. This ambiguity leads to talking past each other about AI's capabilities. We suggest either broadening the definition (e.g., intelligence as the ability to adapt to complex environments) or making it more precise by categorising aspects (such as problem-solving, social, and creative intelligence). Interdisciplinary dialogue among AI researchers, psychologists, and philosophers is critical to develop clear, context-dependent definitions.

6 Artificial general intelligence is not well defined

Artificial General Intelligence (AGI) itself has no universally accepted definition Mitchell [2024a]. ARC-AGI is explicitly a research tool (not a definitive test) focusing on abstraction and reasoning. Critics note that focusing only on "cognitive tasks" ignores embodiment: a system may excel at language but not perform physical tasks ("brain in a vat"). Some have suggested rebranding "artificial intelligence" to "actual intelligence" Mitchell [2019] to reduce anthropomorphism and acknowledge that machines, animals, and humans may each exhibit different forms of intelligence.

7 There may be a spectrum of reasoning in LLMs

Reasoning and intelligence may lie on a spectrum. It is possible that reasoning in LLMs lies somewhere along this spectrum. Reasoning might look very different in machines and humans. The word reasoning comes loaded with many assumptions: it is a metaphor and unfortunately these metaphors are used repeatedly in the field of AI Mitchell [2024b].

8 Alternate View

While this paper argues that intelligence and reasoning in machines and humans are fundamentally different and need distinct definitions, alternative perspectives expand upon this view. An alternative

view asserts that human intelligence and reasoning should remain the gold standard for defining and evaluating machine intelligence. This position is often implicit in benchmarks like ARC, which are designed to test AI systems on tasks that are meant to reflect human cognitive abilities. Proponents argue that the ultimate goal of AI research should be to replicate and surpass human cognitive skills in a way that aligns closely with human definitions of intelligence, as this ensures practical utility in domains like healthcare, education, and governance.

While anthropocentric definitions have practical utility, they risk constraining our understanding of intelligence to human-specific parameters. This ignores the potential for machines to exhibit forms of intelligence that humans cannot. Such a narrow focus may lead to inefficient or misguided approaches to AI development, as it prioritizes mimicking human cognition rather than leveraging the unique strengths of machines. *By redefining intelligence to include machine-specific capabilities, we can create benchmarks and evaluation criteria that account for the distinct ways in which machines operate, fostering innovation beyond anthropocentric constraints*. ARC itself could be adapted to include tasks that explore non-human-centric reasoning, thus broadening the scope of what we consider “intelligent".

The current situation with AI benchmarks is depicted in Figure 1, where a human is asking a seal, a penguin, a dog and a shark to climb up a tree. This ignores the fact that intelligence is task-specific: organisms have evolved over millions of years to solve tasks that are suited for the kind of environment they find themselves in. This frequently involves tradeoffs: a penguin will trade-off flight and efficient locomotion on land for swimming efficiently under water. A human (and other primates) have evolved to walk efficiently on land at the expense of not being able to fly.

Recognizing that intelligence is task-specific and comes in various forms and guises may help us appreciate that animals, humans and machines have different kinds of intelligence, and help us design better benchmarks for machines.

Figure 1: Our current benchmarks for AI are very anthropocentric. It is similar to a situation illustrated here, where a human is asking a seal, a penguin, a dog and a shark to all climb a tree. The implication is that each organism has its own strengths which have been honed through millions of years of evolution to solve certain tasks that are specific to the environment it lives in. Image generated using DALL-E.

9 Discussion

This paper examines the conflict between ARC’s objectives and the employed methods and reflects on whether intelligence and reasoning might ultimately look very different in machines and humans. Some key takeaways are:

1. Goodhart's law: As ARC becomes a target for optimization, its utility as a benchmark may diminish unless new task variations are introduced.

2. Neologisms and new metaphors: We must reconsider the language we use to describe machine behaviours. Terms like "reasoning" and "understanding" may need redefinition to account for the unique characteristics of machine intelligence.

3. Unified theory: A unified approach to meaning and reasoning could help bridge the gap between human cognition and machine abstraction.

4. Reasoning and intelligence may be very different in machines and humans.

At the risk of stating the obvious, let us go over how we humans solve these visual puzzles. We solve them in part using our visual apparatus. Specifically, light reflected from these puzzles impinges on our visual system, where part of the initial processing, such as edge detection, occurs in the eye itself. The remaining processing is carried out in the brain, which integrates this sensory input with a vast amount of prior knowledge, including concepts such as objectness, gravity, etc.

For machines, the processing happens in a fundamentally different way. In the case of large language models, the input is ingested as tokens in a continuous one-dimensional stream, devoid of any direct visual experience. However, if the system is a convolutional neural network, the processing may mimic certain aspects of the human visual system, such as feature detection, but it still lacks the inherent embodiment and experiential learning process of humans.

Machines are also not embodied in the physical world, meaning that they cannot acquire the same training data or learn through interaction and sensory experiences as humans. Their learning is based entirely on the data they are trained on, which is inherently limited compared to the diverse and continuous experiences humans undergo. Hence, to expect that reasoning and intelligence in machines will mirror those of humans when processing the abstraction and reasoning corpus is a fallacy. Differences in sensory processing, embodiment, and prior knowledge result in different pathways of reasoning and intelligence in humans and machines. In other words, there can be many different ways to solve the same problem. *Humans and machines can solve similar problems in fundamentally different ways*.

The computer scientist Edsger Dijkstra, when asked whether computers can think like humans, famously responded with a counter-question: "Do we think submarines swim like fish?" This analogy highlights how the word "swim" is a loaded and problematic metaphor, as it imposes human-centric attributes on non-human entities. Similarly, words like "thinking," "intelligence," and "reasoning" are metaphors we frequently use, and they are what the computer scientist Marvin Minsky referred to as "suitcase words" (terms packed with a range of meanings and assumptions).

For example, while humans have always taken inspiration from birds to achieve flight, human flight looks nothing like avian flight. Airplanes achieve heavier-than-air flight using entirely different principles, such as fixed wings and engines, compared to the flapping of bird wings.

This shows that there can be multiple ways to solve the same problem, each fundamentally different from the other. For instance, the problem of heavier-than-air flight through an atmosphere—has been addressed in ways that differ profoundly between humans and birds. Similarly, the problems of intelligence and reasoning, however they are defined, may also be solved in fundamentally different ways. If broad intelligence is defined as solving problems in a complex novel environment, then machines and humans may come up with fundamentally different ways of solving these problems.

We might need to get used to the idea that machines and humans may have very different kinds of intelligence.

The way machines and humans approach problem-solving can be fundamentally different, yet both can lead to effective solutions. Intelligence and reasoning, however defined, need not be constrained to human-like thought processes. Machines might arrive at solutions that are equally effective, yet completely alien to human cognition.

Instead of insisting that machines must mimic human cognition to be considered intelligent, we should embrace the possibility that intelligence comes in diverse forms, shaped by the constraints and capabilities of the system in which it operates.

References

François Chollet. On the measure of intelligence. 11 2019. doi: 10.48550/arxiv.1911.01547. URL https://arxiv.org/abs/1911.01547v2.

Rich Sutton. The bitter lesson, 2019. URL http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Melanie Mitchell, Alessandro B. Palmarini, and Arseny Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. 11 2023. URL https://arxiv.org/abs/2311.09247v2.

Benjamin Estermann, Luca A Lanzendörfer, Yannick Niedermayr, Roger Wattenhofer, and Eth Zürich. Puzzles: A benchmark for neural algorithmic reasoning. 6 2024. URL https://arxiv.org/abs/2407.00401v1.

Melanie Mitchell. Debates on the nature of artificial general intelligence. *Science*, 383:eado7069, 3 2024a. ISSN 10959203. doi: 10.1126/SCIENCE.ADO7069. URL https://www.science.org/doi/10.1126/science.ado7069.

Melanie Mitchell. *Artificial Intelligence: A Guide for Thinking Humans*. Penguin Publishers, 2019.

Melanie Mitchell. The metaphors of artificial intelligence. *Science*, 386:eadt6140, 11 2024b. ISSN 10959203. doi: 10.1126/SCIENCE.ADT6140. URL https://www.science.org/doi/10.1126/science.adt6140.