A pragmatic metric for artificial general intelligence [DRAFT, WIP]

Why:

Current intelligence definitions are either partial or do not provide an actionable evaluation of intelligence, something tangible which can immediately be used to quantify the potential of a given AGI product. We introduce a new definition, suited for both humans and AIs, which relates intelligence to the ability to perform arbitrary tasks. The main desirable property is the ability of summarising the level of general intelligence to a single number. We also argue that a reliable intelligence measure test cannot be exactly fixed in advance, as any predetermined test can be gamed.

Core Idea:

A single numeric value I, obtained as average with roughly equivalent weights from a small vector of intelligence. The vector is built in 2 steps.

Step 1

The vector is built out of integer values measuring the following specs of the set of tasks:

1. How well the tasks have been solved (accuracy. E.g. 10% accurate = 1 point, 50% = 5 points, 99% = 10 points, 99.9% = 100 …)
2. The tasks to solve are very different from each other (nr of tasks areas covered. E.g language, sequences, etc.)
3. The tasks are complex
4. Few experience (or info) is fed to the system (nr demonstrations, amount of supervised data)
5. Experience is not directly related to the task (overlap tasks - data)
6. Experience is very raw (how much has the data been formatted? without being fed a description of the tasks)
7. Tasks are solved with low resource consumption (time, money, energy, computations, etc.)

Step 2

Given the performance above on a set of tasks, a second set of unknown tasks on different environments is introduced. We measure the percentage in performance change under the following scenarios:

1. New tasks environments are very different. (e.g. the first set of tasks was in videogames, second is in real world.)
2. New tasks cover novel areas.
3. Environment is changed at will dynamically.
4. New tasks areas are added indefinitely.

These factors (< 1) multiply the mean of intelligence vector obtained in step 1.

The key point is that the number and kind of tasks cannot be static. The system must be exposed to new tasks he never trained on to measure its intelligence.

Entrepreneurs, Researchers, Investors, Policy makers can use this metric to quickly assess value.

Benchmarks

Minimum suggested set of tasks to test an agent
Set1:
ARC, bongdard, object manipulation, turing test, raven matrices.
Set2:
Cooking a dish that a person would like, C-test, IMO Grand Challenge

Example
Human and Gato on the same above minimum suggested set of tasks.
Step1 score human: (8, 9, 9, 5, 8, 9, 8)
Step1 score Gato: (6, 9, 9, 3, 2, 3, 4)

Step2 score human: (0.9, 0.8, 0.8, 0.4)
Step2 score Gato: (0.2, 0.3, 0.1, 0.1)

Human I = (8 + 9 + 9 + 5 + 8 + 9 + 8) * 0.9 * 0.8 * 0.8 * 0.4 = 12.9
Gato I = (6 + 9+ 9+ 3+ 2+ 3+ 4) * 0.2 * 0.3 * 0.1 * 0.1 = 0.02

Notice how Human >>> Gato, even though Gato accuracy is just slightly below human.

Perks and Specs:

1. -Task accuracy is just a single component of the evaluation of intelligence.
2. -Intuitive, can be roughly estimated just by looking at the AI architecture and working principles (also due to 1, weak dependence from accuracy).
3. -Based on tasks performances, so independent of the internals of the agent (we are pragmatic and don't care about Chinese room dilemma)
4. -Independent (or weakly dependent) from the set of tasks chosen, assuming that the tasks are comprehensive enough. So different AIs on different sets of tasks can be compared.
5. -Can be applied to already performed tasks, so all past AGIs can be evaluated without further test
6. -Applies to any entity: humans, AIs, superAIs.

Alternatives, doubts and shortcomings:
-Test requires a lot of tasks to be fine grained. What's the minimum set which provides a reliable measure?
-Is the measurement consistent? Does it lead to obvious anomalies?
(check hutter bullet list)
-This measure does not predict intelligence before executing tasks, especially if the architecture is very complicated.
-Even if very intuitive, the aggregate number may be an oversimplification

---

-----—-----

Existing def of intelligence:

Chollet
“the intelligence of a system is a measure of its
skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and
generalization difficulty”
: intelligence is the rate at which a learner turns its experience and priors into new skills at valuable tasks that involve uncertainty and adaptation

Legg Hutter
Intelligence measures an agent’s ability to achieve goals in a wide range of environments.”

“Any system . . . that generates adaptive behaviour to meet goals in a range of environments can be said to be intelligent.” D. Fogel [Fog95]

the ability of a system to act appropriately in an uncertain environment, where appropriate action is that which increases the probability of success, and success is the achievement of behavioral subgoals that support the system’s ultimate goal.” J. S. Albus [Alb91]

“. . . the ability to solve hard problems.” M. Minsky [Min85] “

“Intelligence is the ability to use optimally limited resources – including time – to achieve goals.” R. Kurzweil [Kur00

Achieving complex goals in complex environments” B. Goertzel [Goe06]

Interesting quotes:

Keynote 5.6. Evaluation overfitting in AI: If the problems or instances that represent a task (and their probabilities) are publicly available beforehand, the systems will be specialised for the expected cases instead of the generality of the task. This is known as ‘evalua- tion overfitting’ (Whiteson et al., 2011), ‘method overfitting’ (Falke- nauer, 1998) or ‘clever methods of overfitting’ (Langford, 2005).
(J. Hernández-Orallo)

To reduce the evaluation overfitting problem in a more significant way, the distribution (or subset) that is used to sample instances should not be known in advance (e.g., the "secret generalized methodology", Whiteson et al., 2011). Much better, fresh instances could be generated on the fly.

During the first workshop (Meystel, 2000b), some of the questions were close to those seen during the outset of this book (Panel 1.3), including the definition of a vector of intelligence, the evaluation of intelligence potential and ultimately "whether there exists a universal measure of system intelligence such that the intelligence of a system can be compared independently of the given goals" (Meystel, 2000b). In fact, one of the recurrent debates in the first editions dealt with the distinction between measuring performance and measuring intelligence (Meystel, 2000b). The former was defined in terms of the "vector of performance", a set of indicators of the task the system is designed for. In contrast, the "mysterious vector of intelligence" was "still in limbo", as it should contain "the appropriate degrees of generalization, granularity, and gra- dations of intelligence". The white paper suggested an eclectic set of 25 items for the vector, such as the "number of objects that can be stored", the "ability to assign the optimum depth of associations", the "response time", the "accuracy of the variables", et
Meystel, 2000b

Actually, apart from the areas, they recog- nise six kinds of scenarios: general video-game learning, preschool learning, reading comprehension, story or scene comprehension, school learning and the 'Wozniak Test' (walk into an unfamiliar house and make a cup of coffee). They propose "AGI test suites" such that "the total set of tasks for a scenario must cover all the competency areas" and they must do it in such a way that the tasks are so varied and numerous (or previously unknown) that the "big switch" prob- lem is avoided. Nonetheless, they advocate for several suites, as they "doubt any successful competition could ever cover the full extent of a roadmap such as this" (Adams et al., 2012).

Competency areas in AGI according to Adams et al. (2012).
Perception Attention Planning Actuation Communication Emotion Building/creation
Memory
Social interaction Motivation Reasoning
Learning
Modelling self/other Use of quantities

I-athlon "events" as suggested by Adams et al. (2016). [Courtesy of Sam S. Adams.]

The challenge had two levels. In the first level, the type of IQ tests could be seen beforehand by the AI system pro- grammer. In the second level, however, the types of tests would not have been seen beforehand. Only computers passing the second level "could be said to be truly intelligent" (Detterman, 2011).

"agent is intelligent if and only if it excels at all established, validated tests of intelligence" (Bringsjord, 2011).

Points to Cover

How surprised should I be when a new AGI is announced?

As the number and diversity of tasks increases, the measurement is more reliable.

We provide an intuitive and pragmatic way to evaluate an AGI capability, which can be evaluated by non experts.

Success is based on tasks success, no matter how tasks completion is achieved.

As this definition is not human specific, it can evaluate the intelligence of different entities or combinations thereof. For instance a human with a calculator is effectively more intelligent than one without.

The best way to measure the intelligence of a chess player is to ask him to write a …. (something wildly off from its experience space)

It’s apparent how by this approach no AI system available today gets close to human intelligence.

Tasks need to be unconstrained. Anything goes to achieve them.

How do current AGIs rank?
Gato
GPT-3

Author(s?)

Lorenzo Pieri, …

References:

On the Differences between Human and Machine Intelligence
https://scholar.google.com/citations?view_op=view_citation&hl=en&user=0_Rq68cAAAAJ&sortby=pubdate&citation_for_view=0_Rq68cAAAAJ:C33y2ycGS3YC

On the Measure of Intelligence
https://arxiv.org/abs/1911.01547

S. Legg and M. Hutter, A collection of definitions of intelligence, Frontiers in Artificial Intelligence and applications, 157 (2007),

S. Legg and M. Hutter, Universal intelligence: A definition of machine intelligence, Minds and Machines, 17 (2007), pp. 391-444.
https://arxiv.org/pdf/0712.3329.pdf

P. Wang, On Defining Artificial Intelligence, Journal of Artificial General Intelligence, 10 (2019), pp. 1-37.

J. Hernández-Orallo, The measure of all minds: evaluating natural and artificial intelligence, Cambridge University Press, 2017.

https://www.deepmind.com/publications/a-generalist-agent

Cite the papers in legg, orallo

Random notes:

Then the scale is probably logarithmic.

Benchmarks with humans will be highly useful, but a challenge is to have humans not exposed to previous info.