

Benchmarking Human Intelligence versus AI

DeepMind's Benchmarking of AGI looks like a breakthrough

Tony Czarnecki, Sustensis



London 1/12/2023

For the last two weeks those who follow the events in the area of AI have been preoccupied with the OpenAI saga. We still do not know what was the exact reason for firing and then re-instating Sam Altman as OpenAI's CEO. However, gradually it is becoming clear that the board's rebellion was not just about money but about another breakthrough at OpenAI. It is an apparent discovery of a pathway towards Artificial General Intelligence (AGI). That supposedly happened by combining two OpenAI's teams – the Code Team and the Maths Team. By using a new algorithm called Q* (Q-Star) they may have triggered new capabilities of GPT-4 soon to become GTP-5, alluded by Altman to be a near-AGI.

Almost at the same time something more substantial happened, this time fully revealed, at Google's Deep Mind, which published an article titled: "Levels of AGI: Operationalizing Progress on the Path to AGI" ^[1]. Surprisingly, in this breakthrough article, the authors did not define what AGI is. So, before I continue referring to key concepts in this important article, let me fill in the gap, since this term has not been universally agreed yet, by proposing the following definition of AGI:

Artificial General Intelligence is a self-learning, superior to humans' intelligence, capable of solving any task far better than any human.

In simple terms, we will know that AGI has arrived when AI will be smarter than any human in any area. Perhaps the best way to illustrate it, might be to

compare the AI's intelligence with that of the most able humans (that's what the authors' article also suggest). It is possible that humans might still excel AI in some capabilities and skills because of the biological nature of our bodies. For example, a humanoid will not be able to feed a baby as a mother does. Conversely, AI can already do things, which humans will never be able even closely to match, like speaking fluently in over 100 languages.

Although the authors do not provide their own definition of AGI, they propose, what they call, 'operationalizable definition'. It is based on 9 examples of AGI definitions, provided by other authors. The authors of the article use them to provide 6 properties and commonalities of AGI as the basis for a certain focus (direction) of AI research:

1. Focus on **Capabilities**, not Processes
2. Focus on **Generality** and Performance
3. Focus on **Cognitive** and Metacognitive Tasks
4. Focus on **Potential**, not Deployment
5. Focus on **Ecological Validity**
6. Focus on the **Path to AGI**, not a Single Endpoint

That helps them to define 6 levels of AI Competence, like for the self-driving cars:

- Level 0 - **No AI**
- Level 1 - **Emerging** – *equal, or slightly better than an unskilled human*
- Level 2 - **Competent** - *at least 50th percentile of skilled adults*
- Level 3 - **Expert** - *at least 90th percentile of skilled adults*
- Level 4 - **Virtuoso** - *at least 99th percentile of skilled adults*
- Level 5 - **Superhuman** - *outperforms 100% of humans*

If the AI sector accepts this AI ontology, as happened for self-driving cars, and adopts these competencies as the guidance for developing AGI, then this may become an effective starting point for determining the current competency level of the maturing AGI. That is why this article is so important. It has also entered the uncharted waters of defining intelligence, necessary to establish a more precise meaning of human-level intelligence.

What does 'human level intelligence' really mean?

This is the area, which I have considered very important for some time since this might make AI development control more effective. But it is also the area of several unknowns such as consciousness and cognition. Therefore, anyone venturing to debate intelligence, risks being misunderstood for putting forward ideas or concepts, which have no sound foundation in reality. Nevertheless, it may be helpful to use comparisons between human and AI intelligence at least to approximate how close AI is from being superior to humans, especially that even some AI researchers consider it a new type of technology. **But AI is foremost a new, inorganic intelligence.** That new intelligence may achieve its goals and solve problems differently than we do, being smarter than any human, even if it does not tick all the boxes on the human intelligence definition, e.g. abstraction.

When comparing the AI's and human intelligence most authors use a definition which clarifies what it is from a human perspective. For example, Encyclopaedia Britannica defines it as follows: '*Human intelligence - mental quality that consists of the abilities to learn from experience, adapt to new situations, understand and handle abstract concepts, and use knowledge to manipulate one's environment*'^[7].

However, for a more objective comparison, we should consider intelligence from the perspective of the Universe, in which there may be different forms of intelligence, of which biological intelligence may be just one. This is similar to Ray Kurzweil's thinking when he said in 2023 'The universe has been set up exquisitely enough to have intelligence. There are intelligent entities like us who can contemplate the universe and develop models about it, which is interesting. Intelligence is, in fact, a powerful force and we can see that its power is going to grow not linearly but exponentially and will ultimately be powerful enough to change the destiny of the universe'^[8]. Thus, my definition of intelligence in the context of the Universe is as follows:

'Intelligence is an attribute of an organic or inorganic system, which intentionally changes its environment to achieve its goals using minimum amount of energy'.

The condition of a minimum amount of energy is necessary for the intelligent being to evolve and be even more intelligent. From that point of view, pansychism at a macrophenomenal level offers some explanation of how intelligence, as an attribute of a generic mind, may actually work. In summary, AI's and humans' intelligence should be compared from the universal perspective rather than from a strictly anthropic point of view.

Now, coming back to DeepMind's article, since they have not defined what intelligence is, it is difficult to see how the Competence (Autonomy) Level is linked to intelligence. To do that, I have applied the Multiple Intelligence Theory created by Howard Gardner, an MIT professor of psychology at Harvard University. It challenges traditional beliefs in the fields of education and cognitive science. For example, the American Psychological Association defines intelligence as follows: '*Intelligence is the ability to derive information, learn from experience, adapt to the environment, understand, and correctly utilize thought and reason.*' The key word in this definition is ***understand***. According to that traditional definition, intelligence is a uniform cognitive capacity people are born with. This capacity can be easily measured by reasonably simple tests. But according to Gardner, intelligence is:

- The ability to create an effective product or offer a service that is valued in a culture,
- A set of skills that make it possible for a person to solve problems in life,
- The potential for finding or creating solutions for problems, which involves gathering new knowledge,

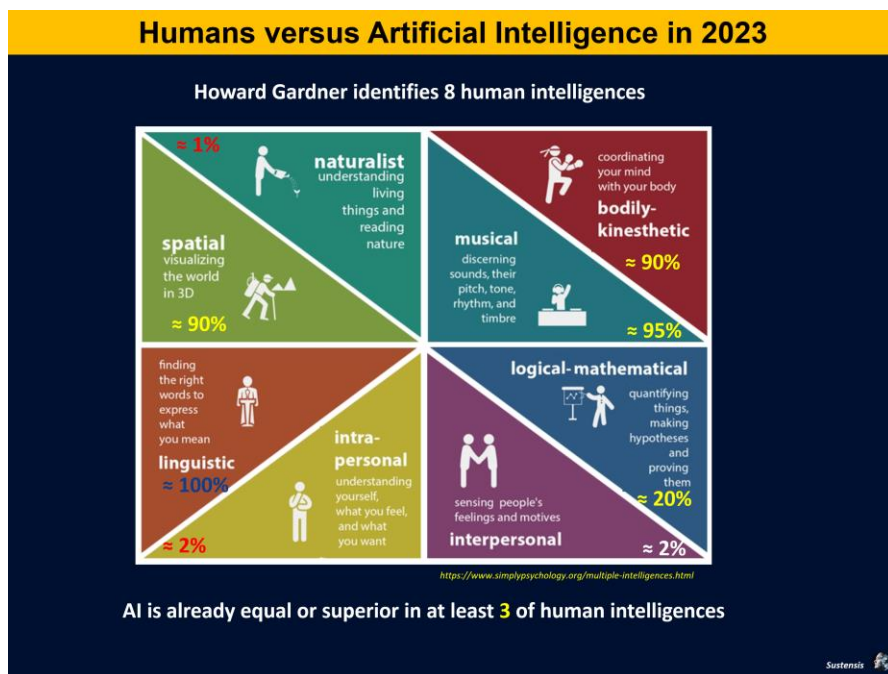
In addition, Gardner postulates that:

- All human beings possess all intelligences in ***varying*** amount,
- Each person has a ***different*** intellectual composition,
- These intelligences are located in different areas of the brain and can either work independently or together,
- These intelligences may define the human species,
- Multiple intelligences can be nurtured and strengthened, or ignored and weakened,
- Each individual has eight intelligences (and maybe more to be discovered)^[9].

Unlike the majority of theories of intelligence, Gardner's theory of Multiple Intelligences proposes a differentiation of human intelligence into specific modalities of intelligence, rather than defining intelligence as a single, general ability. The theory has been criticized by mainstream psychology for lack of empirical evidence, and its dependence on subjective judgement^[10].

However, I would suggest the opposite. The arguments that a human brain is unlikely to function using Gardner's multiple intelligences, are precisely the reason why his theory is more useful for comparison with the AI's intelligence. As can already be seen, AI's intelligence is, or at least it may be, of a different kind than human's.

Howard Gardner identifies 8 human intelligences [11]: Linguistic, Logical/Mathematical, Spatial, Bodily-Kinaesthetic, Musical, Interpersonal, Intrapersonal, and Naturalist. Skills are mainly about *doing*, whereas intelligence is more about contextual *understanding*. Therefore, in my comparison some of AI's 'intelligences', are simply *skills or competencies*, which are needed to perform a task that requires intelligence.



How human intelligence compares against Artificial Intelligence today.

I have estimated how well AI currently matches human intelligence in each of the eight intelligences based on the AI's skills, and NOT on AI's intelligence viewed from a human perspective. Looking at individual skills, rather than intelligence type as a whole, is more relevant approach because it compares the real impact of AI on us and the environment, i.e., on what really matters, at least from the point of view of AI development control.

I should also clarify which level of human intelligence I am comparing: the most intelligent people, or an average human. DeepMind suggests that AGI should have superhuman capabilities, i.e., surpassing any human in competence and intelligence. I also assume we are assessing AI against most intelligent people, e.g., when evaluating postgraduate exam results and, each comparison made only for a particular type of intelligence. For example, in translation there is no human capable of speaking and fluently translating 100 languages. However, when it comes to translating poetry, humans still excel due to their understanding of language nuances. The question then arises: which superiority is more significant in achieving life goals, i.e., being smarter and having a better chance of surviving in a dangerous situation? It's the ability to translate 100 languages simultaneously, i.e. to communicate, the area in which, AI is already vastly superior. It may be only slightly behind top human translators in non-technological areas. Therefore, the score of 90% is not an exaggeration.

My other assessments follow a similar approach, verifying what is most useful for achieving life goals. From that perspective, in four of the intelligences, Bodily-Kinaesthetic, Linguistic, Musical and Spatial, AI already equals or exceeds top-performing humans. For example, in bodily kinaesthetic, some robots like Boston Dynamics' Atlas, can perform nearly acrobatic jumps, comparable with top human gymnasts. If we were considering just that skill alone, AI achieves about 90% level.

However, in maths, when performing error free complex calculations, AI is still far behind humans. It can prove theorems, but is not capable yet of creating its own, because that would require genuine creativity, which it does not have. Therefore, in that area, AI is perhaps only at about 20% level.















In music, including composition, art (painting) or literature, it is quite often impossible to distinguish the products created by AI from those created by humans. Therefore 80% score seems to be reasonable.

Similarly in spatial visualization tasks, such as object recognition in images, or 3D object reconstruction from images, AI systems have achieved high levels of accuracy. For instance, in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), the winning deep learning models achieved top-5 error rates of less than 5%, which is lower than the human error rate of about 5-10%. In 3D object reconstruction, AI systems can reconstruct detailed 3D models of objects from multiple images with a high level of accuracy.^[12] This data is about 10 years old. Today, the error rate

should be even lower. Therefore, a score of 90% seems reasonable in this category, although of course there are areas where humans still excel AI, especially in creativity, which is unmeasurable. Humans are still immensely superior in Interpersonal, Intrapersonal (understanding yourself, emotions etc), and in Naturalist areas. These are the areas closely related to cognition.

In addition to comparing human's intelligence against AI's intelligence at an aggregate level, I have compared it directly at a particular skill level. Here, a very fast progress has been noted in sensory processing, a crucial component for developing AI's cognitive capabilities [6]. AI's perception skills, like taste, smell or touch have been available for at least two years. They are far more accurate than human senses but need to be integrated, like other modes: video, sound, verbal communication etc.

AGI may emerge and reach human level intelligence by 2030

SKILLS/CAPABILITIES		SENSORY PERCEPTION	
<p>2019 - Composer - AIVA - 100%</p>  <p>2021 - Artist - AI-Da - 100%</p>  <p>2021 - Author - GPT-3 - 100%</p>  <p>2022 - Research - Iris' Research Workspace - 90%</p>  <p>2022 - 'Labourer' - Boston Dynamics' Handle - 100%</p> 	<p>2022-11-30 - First 'human-like' Content Generator - OpenAI - 50%</p>  <p>2023-03-01 - Kosmos-1 First multi-modal 'human-like' Content Generator - OpenAI - 70%</p>  <p>2023-03-07 - PaLM-E First generalist robot acting on human commands - Google - 60%</p>  <p>2023-11-08 GPT-4 Turbo: competent conversation with live Internet access and integrated APIs - 90%</p> 	<p>2021 - Image recognition - Clearview - 100%</p>  <p>2022 - Touching - 50%</p>  <p>2022 - Emotions (showing, not feeling) - AMICA - 90%</p>  <p>2022 - Smelling - Aryballe's NoSE Advance - 100%</p>  <p>2022 - Audio incl. Translation - Google Translator - 90%</p> 	<p>2022 - Emotions (showing, not feeling) - AMICA - 90%</p> <p>2022 - Smelling - Aryballe's NoSE Advance - 100%</p> <p>2022 - Audio incl. Translation - Google Translator - 90%</p>

©Sustensis

We have made a relatively modest progress in those areas where AI is not yet at a human level, e.g., a Generalist Robot (or Universal Assistant). That may change soon, since OpenAI and DeepMind, the two leading AI research organisations, are currently changing the approach from relying mainly on the two big planks in AI – Large Language Models (LLM) and Deep Learning. They are shifting towards a more general-purpose, cognitive

approach, to create systems that can reason and learn across a wide range of domains, rather than just excel at specific tasks or problems. If the story about Q* is true then it might confirm why resolving this problem in principle may have such a fundamental significance.

It is possible that the problem of cognition may solve itself when humanoid robots such as Ameca or Optimus, arrive in larger numbers operating in real environment as humans' co-workers. Elon Musk thinks his Tesla cars are the most cognitive AI systems, which must be aware of thousands of situations in a second. Therefore, once that software is uploaded with some modifications to an Optimus humanoid robot, cognition, including self-awareness, may just happen spontaneously.

The approach taken by the authors of the paper offers *'a more nuanced way to define our progress toward AGI by considering generality (either Narrow or General) in tandem with five levels of performance (Emerging, Competent, Expert, Virtuoso, and Superhuman)*.

Therefore, they correlated these Levels of AI competence with Levels of Autonomy as follows:

- **Autonomy Level 0:** No AI - human does everything
- **Autonomy Level 1:** AI as a Tool - human fully controls task and uses AI to automate mundane sub-tasks
- **Autonomy Level 2:** AI as a Consultant AI takes on a substantive role, but only when invoked by a human
- **Autonomy Level 3:** AI as a Collaborator co-equal human-AI collaboration; interactive coordination of goals & tasks
- **Autonomy Level 4:** AI as an Expert AI drives interaction human provides guidance & feedback or performs subtasks
- **Autonomy Level 5:** AI as an Agent fully autonomous AI

The key message is that it's the level of AI competence that will ultimately impact the scope of its autonomy. Once it has achieved level 5 in all competencies, it will become AGI.

From our point of view, **the most important conclusion is that AGI will not emerge instantaneously but rather progressively both in the scope and level of competencies.** That is significant for two reasons. First of all, it is directly related to the ability of controlling AI development. If AI is

developed by continuously measuring its benchmark competence level and indirectly its level of autonomy, then the AI researchers and developers will have much greater scope of control and plenty of warning of emerging AGI.

On the other hand, because of AI's self-learning capabilities, once it acquires some cognitive functions, it might increase its competence level and scope very quickly, perhaps in weeks, and become AGI. The best example is the pace of GPT product line improvement. Version 3.0 had a maximum 1 page of context memory and quite a laborious process of tuning AI, with no attachments and no tools. It took about 2 years for version 3.5. to arrive in the shape of ChatGPT. No direct attachment and no API tools could be used. Within a few months, GPT 4.0 has arrived, which had 3 pages of context memory, allowed to attach a file and use some API tools, such as Excel or DALL-E 2. And then within a few months, version GPT 4.5 Turbo emerged, which now enables about 250 pages of context memory, multiple attachments, and use of APIs in the background.

The release of this paper has quite significantly increased the chances of AGI emerging in the next few years because developers will have to deal with fewer unknowns related to accidentally producing AGI, before global mechanisms of control have been established. The frontier models will be measured and compared against a benchmark once it is globally accepted. It will become a valuable tool for assessing how close AI is to become AGI and what risk the most advanced model may pose to humans. The monitoring of that process should be the key task of the Global AI Safety Institute set up at Bletchley Park Summit in November 2023. However, developers will still have to be prepared for uncomfortable surprises of a sudden increase of near AGI's capabilities, where none were expected. That is what happened when ChatGPT was released.

In summary, this breakthrough paper shows that AGI is unlikely to be reached in one instant. Instead, it will arrive in stages, defined by Autonomy levels (capabilities or competencies). Measuring the AI's Autonomy level, i.e. its independence in defining and carrying out its goals, may give us a clear warning when AGI may emerge. Although it does not remove entirely the overall problem of controlling AGI, it may in some way delay or even stop development of the most advanced AI, when it is still at a lower level of Autonomy. This would give us more time for aligning the AGI's goals and behaviour with humans' values and preferences. Therefore, the approach the article proposes, presents rather an optimistic perspective for a more effective AI development control, at least for the next few years.

For all that to work, there would need to be a single, [Global AI Development Centre](#), under the supervision of an international organization with powers similar to International Atomic Energy Authority (IAEA).