proceedings

Levels of Intelligence in Artificial and Biological Systems †

Howard Schneider[1]*

[1] Sheppard Clinic North
* Correspondence: hschneidermd@alum.mit.edu

† Presented at the 2021 Summit of the International Society for the Study of Information (IS4SI): Information in Biologically Inspired Cognitive Architectures (BICA) Vienna, Austria, September 12 – 19, 2021

Abstract: A two-dimensional rating scale is proposed to measure more readily what we consider the intelligence of an artificial intelligence system or a natural cognitive system. In the proposed rating scale, one axis measures the causal abilities of the system, while the other axis measures the system's data processing abilities, which is called its benchmark at that level of causality. The benchmark value is the common logarithm of the raw values of the data processing parameters being measured. For example, a given AI system's intelligence value could be measured as, Level 3 Benchmark 7, where the Level 3 refers to its causal level, and the Benchmark 7 refers to the data it can access and the compute it can use. Causal levels vary from 0 (no or few organized associations) to 7 (fully cause-and-effect mechanisms). Benchmarks generally are between 0 and 10 but depend on a reference value for that level of causality. This two-dimensional rating scale for intelligence allows the relatively simple application of the scale to an artificial intelligence (or natural) system, and easy communication of the results to another party.

Keywords: Intelligence tests; Machine Intelligence; Artificial General Intelligence; Artificial Intelligence; Causality

1. Introduction—The Need to Classify and Measure Levels of Intelligence

At the time of writing, the field of artificial intelligence is still unfortunately filled with much exaggeration compared to its actual achievements [1, 2, 3, 4, 5, 6]. The definition of what is meant by artificial intelligence can be a broad one—from classification of images in a well-defined distribution (e.g., deep learning recognition of faces), to using various data science principles (e.g., automating aspects of statistical analysis of a large body of data), to playing a games such as Go or chess at an expert level [7], to writing convincing stories (e.g., Generative Pre-Trained Transformer 3 (GPT-3) [8]), to helping clinicians making better medical decisions (e.g., Shortliffe [6]). All these applications of technology have tremendous potential but also, at the time of this writing, significant limitations. For example, while GPT-3 can write essays that appear at a human level, this may only apply to limited portions of the essay viewed without much context. If GPT-3 is asked what is heavier, for example, a pencil or a toaster oven, it does not know the answer, let alone understand the question [9]. There is a need to classify and measure a "level" of some quantity which reflects what would be a "reasonable" assessment of a system's "intelligence."

Citation: Schneider, H., Levels of Intelligence in Artificial and Biological Systems, …..

Published: date

Much work has already been done on attempting to both define and measure artificial intelligence. For example, Legg and Hutter [10] provide a mathematical formulation of a measure of machine intelligence (1). It gives the expected performance Υ of agent π over the algorithmic probability distribution of the space of environments $2^{-K(\mu)}$ times the value function V of agent π operating in environment μ (where μ is one environment in the set of E all environments that could exist), i.e., intelligence is measured in terms of an agent's "ability to achieve goals in a wide range of environments." $\Upsilon(\pi)$ effectively represents Legg and Hutter's "universal intelligence" of agent π. However, K is the Kolmogorov complexity function, which is not readily computable for real-world environments. Thus, while this equation can theoretically give a universal measure of intelligence for many types of intelligent machines, it really is not practical to use. Goertzel [11] criticizes this approach in ignoring the complexities of real-world cognitive architectures. For example, many cognitive animals may not be able to function well in set of all possible environments, and they do not necessary spend every moment trying to obtain some defined external reward. Of interest, Schmidhuber and colleagues [12] provide some practical workarounds to the incomputability of Legg and Hutter's universal intelligence, with the use of games not only as benchmarks for narrow AI but extending them as measures of more general AI systems.

$$\Upsilon(\pi) := \Sigma_{\mu \epsilon E}\, 2^{-K(\mu)}\, V^{\pi}{}_{\mu} \qquad (1)$$

Chollet [13] notes that to measure intelligence in terms of equation (1) there is the need for an intelligent agent to be able to perform many different tasks that the many different environments would require. Thus, there is the need for generalization whereby the agent would have to be able to learn new skills to successfully perform the many possible tasks, that often may be different than previously learned tasks. Chollet thus proposes the Abstraction and Reasoning Corpus (ARC) to measure generalized machine intelligence. The tasks in ARC are similar to Raven's Progressive Matrices found on human intelligence tests, although with greater task diversity. While ARC may be useful to benchmark future advanced AI systems, it may be less useful to characterize current AI systems.

There have been many attempts to characterize what is AI or what is Artificial General Intelligence (AGI). For example, Adams and colleagues [14] describe a large variety of characteristics for AGI environments, tasks, agents, and architectures. Wang [15] attempts to carefully define what is an artificial intelligence, and proposes "adaptation with insufficient knowledge and resources." Rosenbloom and colleagues [16] describe characterizing AI systems and cognitive architectures in terms of basic dichotomies, i.e., is the system symbolic versus sub-symbolic, symmetric versus asymmetric, and combinatory versus non-combinatory. However, like the other descriptions and benchmarks proposed above, these approaches are not all that useful for describing the measure of intelligence an artificial intelligence system has. There is the need to better and more conveniently classify and measure levels of intelligence in AI and cognitive systems.

2. Measuring Levels of Intelligence in Natural and Artificial Intelligence Systems

In discussing and comparing different autonomous driving systems, Standard J3016 of the Society of Automotive Engineers has proven useful [17]. Level 0 is no automation (i.e., the human driver is fully responsible for steering, accelerating, and decelerating) while Level 5 is full automation (i.e., the machine intelligence can handle all aspects of steering, accelerating, and decelerating under all conditions where a human drive could drive). The levels in between represent partial aspects of driving automation. In Level 2 driving automation, for example, the machine intelligence can steer and accelerate/decelerate under certain driving conditions with the human expected to monitor and takeover when necessary, while in Level 4 automation the human is not required to monitor the machine intelligence, although the latter may not be capable of handling every driving situation a human could (and in such a less common situation the machine intelligence may automatically slow and pull the car over to the side of the road).

Without having to get into the details of whether a self-driving car uses this or that algorithm, or this or that hardware enhancement, and so on, the driving levels of automation allow us to specify in a common language, the rough abilities of a particular autonomous car. For example, if we say that "self-driving" cars from manufacturer A operate at Level 4 autonomy while "self-driving" cars from manufacturer B operate at Level 2 autonomy, we quickly obtain knowledge about the approximate self-driving abilities of the brand A automobile versus the brand B one.

Similar levels of, for example, "intelligence" to characterize natural and artificial cognitive systems would be beneficial. However, five simple levels representing the level of "intelligence" a machine intelligence or cognitive system possesses, may fail to deliver meaningful results. As Goertzel [11] above notes, assigning a universal intelligence value based on some defined external reward or other criterion, ends up ignoring complexities of the real-world—an intelligent system or animal may function very well in a particular environment for a limited lifetime, and trying to assign some universal intelligence value as such may simply not be that useful.

Adams and colleagues [14] describe eight characteristics of AGI-capable agents, including the tasks they should perform and the environments they should operate in:

1. diverse, interacting, and structured objects in the environment;
2. environment is open—not a fixed set of objects and events;
3. different timescales for regularities in tasks;
4. interactions with other agents;
5. tasks can be novel;
6. complex interactions between the agent and the environment and tasks;
7. the agent does not have unlimited compute;
8. the agent exists over a "long-term" and continuously.

While we could create an eight-dimensional "intelligence" rating system based on the criteria above, it would not meet our goal of having a relatively easy to understand and communicate system similar to Standard J3016 of the Society of Automotive Engineers which allows straightforward assessment and communication concerning self-driving capabilities.

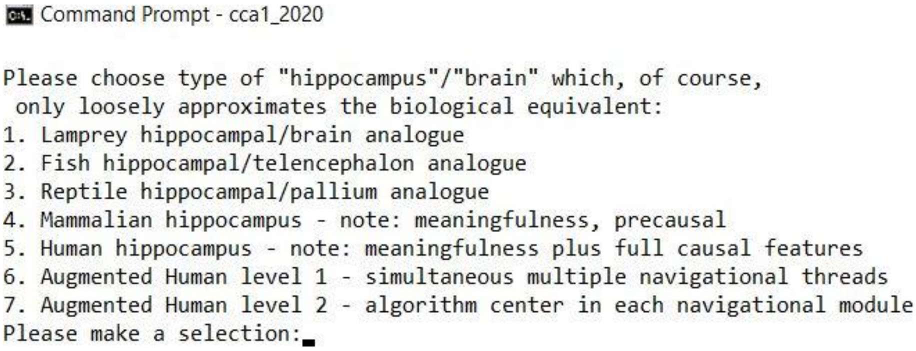
To reduce the dimensionality of the above list, we replace it with a two-dimensional system which we propose can actually meet the same functionality as the above list. Much as Legg and Hutter [10], Chollet [13], and other researchers discussed above, try to essentially measure adaptability of systems as a measure of their intelligence, the causal abilities of intelligent systems greatly affect this adaptability. Pearl [18] describes a three-level causal hierarchy. As will be shown in the next section, while we don't use Pearl's levels exactly, we base one dimension of our intelligence rating system on it. For the second dimension in our intelligence rating system, we use a rating of the data processing power the system has at a particular level. For example, two intelligent systems may be on one axis at Pearl's "Association Level" but one system may only be able to perform a few associations while the other system may be able to perform billions of associations. The latter system is considered to have more intelligence than the former.

3. A Two-Dimensional Rating Scale for Levels of Intelligence in AI and Natural Systems

As discussed above, we approach the measurement of intelligence in artificial and natural systems with a two-dimensional rating scale. One axis measures the causal abilities of the system, albeit at a more nuanced level than Pearl's ladder, while the other axis essentially benchmarks the system's data processing abilities at that level of causality.

For the causal axis, we consider the vertebral brain, where there may be a variety of pre-causal levels of functionality, i.e., not actual cause-and-effect action, but more than associations [19, 20, 21]. As shown in Schneider [21] the use of navigation maps for a system of intelligence can allow cognitive architectures (as well as animal brains) which do not have full causal abilities but have behavior which seems to exhibit superficially cause and effect characteristics. Indeed, in Schneider's Causal Cognitive Architecture [21] the user can select an "animal-like" level in which to run the simulation. The selection menu from the simulation is shown in Figure 1.

.



```
Command Prompt - cca1_2020

Please choose type of "hippocampus"/"brain" which, of course,
 only loosely approximates the biological equivalent:
1. Lamprey hippocampal/brain analogue
2. Fish hippocampal/telencephalon analogue
3. Reptile hippocampal/pallium analogue
4. Mammalian hippocampus - note: meaningfulness, precausal
5. Human hippocampus - note: meaningfulness plus full causal features
6. Augmented Human level 1 - simultaneous multiple navigational threads
7. Augmented Human level 2 - algorithm center in each navigational module
Please make a selection:
```

Figure 1. User can select associative, pre-causal or causal features of the Causal Cognitive Architecture simulation

The proposed two-dimensional rating system for intelligence in artificial (or natural) systems is shown in Table 1. The benchmark value of the artificial example defines quantitatively the second axis of the rating system. The benchmark value is the common logarithm of the raw benchmark data processing value, i.e., as shown in (2) below.

$$\text{benchmark\_value} = \log_{10}(\text{raw\_data\_processing\_value}) \qquad (2)$$

The benchmark value shown in Table 1 can be used to calculate benchmark values for other systems. Thus, at the Level 1 row, we see that for the given benchmark $\log(10^9/k) = 5$, thus $k = 10^4$. If, for example, we had a data lookup table capable of similar processing but with 100,000 entries, then the benchmark value would be $\log(10^5 /k)$ or 1 (and thus we would describe the intelligence of this system as Level 1 Benchmark 1 intelligence).

A full discussion of the natural examples in Table 1 is beyond the scope of this paper. For example, in the Level 0 row, even if there are only some weak associations present, if there are enough of them, interesting behaviors can often emerge [22]. As well, in comparing vertebrate neuroanatomy of different species, Butler and Hodos [23] caution that the concept of *scala naturae* arranging vertebrates in a sequence from "fish to reptile to… human" may seem intuitively correct, but all should be considered as successful in adapting to their environments.

The two-dimensional rating scale allows useful classification of AI systems and easy communication of this information. As such, one can abstract away many of algorithmic methods and hardware implementations of an AI system (as well as much hype, even if it is unintended). For example, a developer may have created a new AI system with this and that algorithm and data. However, we can simply say that it is an AI system, for example, with Level 2 Benchmark 4 intelligence, and this information can be conveyed easily to another party.

If different intelligent systems are combined, then the system with highest level should be used. For example, what is the rating of a human working with a powerful search engine? The causal level of the human is higher so the level rating is at Level 6, but the search engine increases the data processing power of the human, so the benchmark rating rises to 6, so the overall rating of the intelligence of this system (i.e., human using powerful search engine) would be Level 6 Benchmark 6 intelligence.

Table 1. A Two-Dimensional Rating Scale for Levels of Intelligence in AI and Natural Systems

Note: benchmark_value = log$_{10}$(raw_data_processing_value)

Level of Intelligence	**Natural Example**	**Artificial Example**	**Benchmark Value (of Artificial Example)** **(log (raw data processing))**
Level 0 – No or Few Organized Associations	Spores blowing in the wind	Digital clock	2
Level 1 – Reflexive Associations	Bacterial chemotaxis	Data lookup table with one billion entries	5
Level 2 – Complex Associations	Fish simple behaviors	Convolutional Neural Network-- can recognize one million faces	5
Level 3 – Complex Associations with Specialized Processing Centers	Fish complex behaviors	Generative Pre-Trained Transformer Neural Network with 175 billion parameters	7
Level 4 – Complex Associations plus some Pre-Causal Associations	Reptile	Experimental [e.g., Causal Cognitive Architecture [21]	1
Level 5 – Fully Pre-Causal Associations	Mammal	Experimental [e.g., Causal Cognitive Architecture [21]	1
Level 6 – Pre-Causal plus some Cause-and-Effect Logic	Human	not available	Natural Example of human := 5
Level 7 – Fully Cause-and-Effect Mechanisms	not available	not available	not available

4. Conclusion and Further Work

Above we reviewed some of the literature that attempts to define or measure machine intelligence. We then proposed a two-dimensional rating scale to measure more readily what we consider the intelligence of an artificial intelligence system or a natural cognitive system. In the proposed rating scale, one axis measures the causal abilities of the system, while the other axis measures the system's data processing abilities, which is called its benchmark at that level of causality. The benchmark value is the common logarithm of the raw values of the data processing parameters being measured. For example, a given AI system's intelligence value could be measured as, for example, Level 3 Benchmark 7, where the Level 3 refers to its causal level, and the Benchmark 7 refers to the data it can access and the compute it can use. Causal levels vary from 0 (no or few organized associations) to 7 (fully cause-and-effect mechanisms). Benchmarks generally are between 0 and 10 but depend on a reference value for that level of causality. This two-dimensional rating scale for intelligence allows the relatively simple application of the scale to an artificial intelligence (or natural) system, and easy communication of the results to another party.

Future work includes testing this two-dimensional rating system on additional AI and natural systems, and measuring its strengths and weaknesses on different AI and natural systems.

Funding: This research received no external funding

Informed Consent Statement: Not applicable

Data Availability Statement: Please contact hschneidermd@alum.mit.edu

Conflicts of Interest: The author declares no conflict of interest.

References

1. Darwiche, A. (2017). Human-Level Intelligence or Animal-Like Abilities? *Communications of the ACM* 61(10). doi: 10.1145/3271625

2. Horgan, J. (2020). Will Artificial Intelligence Ever Live Up to Its Hype? *Scientific American* Dec 4. Retrieved from: https://www.scientificamerican.com/article/will-artificial-intelligence-ever-live-up-to-its-hype/

3. Hutson, M. (2018). Has artificial intelligence become alchemy? *Science* 478(6388):478. doi: 10.1126/science.360.6388.478

4. Hutson, M. (2020). Core progress in AI has stalled in some fields. *Science* 368(6494):927. doi: 10.1126/science.368.6494.927

5. Perry, T.S. (2021). Andrew Ng X-Rays the AI Hype. *IEEE Spectrum*, May 3. Retrieved from: https://spectrum.ieee.org/view-from-the-valley/artificial-intelligence/machine-learning/andrew-ng-xrays-the-ai-hype

6. Shortliffe E. H. (2019). Artificial Intelligence in Medicine: Weighing the Accomplishments, Hype, and Promise. *Yearbook of medical informatics*, *28*(1), 257–262. https://doi.org/10.1055/s-0039-1677891

7. Schrittwieser, J., Antonoglou, I., Hubert, T. *et al.* (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588**:**604–609. https://doi.org/10.1038/s41586-020-03051-4

8. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

9. Lacker, K. (2020). Giving GPT-3 a Turing Test. *Kevin Lacker's Blog*. July 6. Retrieved from: https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

10. Legg, S., Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. arXiv: 0712.3329

11. Goertzel, B. (2010). Toward a Formal Characterization of Real-World General Intelligence. *Proceedings of the 3rd Conference on Artificial Intelligence*. doi: org/10.2991/agi.2010.17

12. Schaul, T., Togelius, J., Schmidhuber, J. (2011). Measuring Intelligence through Games. arXiv:1109.1314

13. Chollet, F. (2019). On the Measure of Intelligence. arXiv:1911.01547

14. Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S., Sowa, J. (2012). Mapping the Landscape of Human-Level Artificial General Intelligence. *AI Magazine*. 33:25-42. 10.1609/aimag.v33i1.2322.

15. Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*. 10(2):1-37. doi: 10.2478/jagi-2019-0002

16. Rosenbloom, P., Joshi, H., and Ustun, V. (2019). (Sub)Symbolic x (A)Symmetric x (Non)Combinatory: A Map AI Approaches to Spanning Symbolic/Statistical to Neural/ML. In: Cox, M.T. (ed) Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems.

17. Hopkins, D., Schwanen, T. (2021). Talking about automated vehicles: What do levels of automation do? *Technology in Society*. Vol 64. doi: 10.1016/j.techsoc.2020.101488

18. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM* 62(3):54-60. doi: 10.1145/3241036

19. Schneider, H. (2018). Meaningful-Based Cognitive Architecture. *Procedia Computer Science BICA 2018*, ed Samsonovich, A.V., 145:471-480.

20. Schneider, H. (2020). The Meaningful-Based Cognitive Architecture Model of Schizophrenia. *Cognitive Systems Research* 59:73-90. doi.org/10.1016/j.cogsys.2019.09.01

21. Schneider, H. (2021). Causal cognitive architecture 1: Integration of connectionist elements into a navigation-based framework. *Cognitive Systems Research* 66:67-81. doi.org/10.1016/j.cogsys.2020.10.021

22. Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press, NY, USA.

23. Butler, A.B. and Hodos, W. (1996). *Comparative Vertebrate Neuroanatomy*. Wiley-Liss, NY, NY.