Reaching for the intangible

Jacob Valdez

September 12 2021

Exactly how intelligent would you say humans are? With Hunter and Legg's definition of intelligence as the ability to achieve a wide range of goals in many environments, it's safe to say that humans are highly intelligent in comparison to all other observed goal-seeking systems. In fact, it is extremely difficult to identify goals they cannot seek within the framework of our universe. However, unreachable goals do 'exist' as identified by Godel's incompleteness theorems and Turing's halting problem. Those formal proofs established the unprovability, undecidability, and intractability of their decade- and century-long problem domains, and their eager or reluctant acceptance optimized the ambitions of the mathematical and computational science research that followed.

I see certain fields of AI today reaching for a similarly intangible goal: human-level artificial intelligence. You see, 'human-level' AI research often begins with the speculation: "I see that human intelligence uniquely does this or that, so if I make an AI system with those features, it must be 'human-level' artificial intelligence." Consider three examples of this thinking in action:

- Turing [12] proposed the Imitation game as a discriminative test of humanly-indistinguishable machine "thinking".

  > 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this [between a machine and a woman] as he does when the game is played between a man and a woman? [...] The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include.

- Nilsson [9] proposes the employment test as identifying a the fractional degree of progress towards human-level AI:

  > To pass the employment test, AI programs must be able to perform the jobs ordinarily performed by humans. Progress toward human-level AI could then be measured by the fraction of these jobs that can be acceptably performed by machines

- You [10] propose the language acquisition test:

> We conjecture that learning from others' experience with the language is the essential characteristic that distinguishes human intelligence from the rest. Humans can update the action-value function with the verbal description as if they experience states, actions, and corresponding rewards sequences firsthand.

While these theses provide concrete measures which are useful feedback signals to compare AI, I find them insufficient to fully define the meaning of 'human-level' intelligence. It is now common to read statements from gpt3 and other large language models that pass our subjective 'Turing test'. The employment test is probably the most general of the above three measures of human-level AI since it intrinsically demands few-shot on-the-job learning, but AI evolution is a slow feedback signal and better suited as an auxiliary dev metric than a primary optimization objective. We would may end up discovering the divergence between working-human intelligence and unemployed human intelligence by the time that benchmark approaches 100%. Finally, the language acquisition test is trivially solved by only reducing the parametric complexity of the "action-value function". Learned optimizers have already been shown capable of optimizing themselves [7, 8], and in the language of reinforcement learning, you might say that they update their own action-value function. With various definitions of "language", multi-agent reinforcement learning has also identified language emergence, acquisition, and guidance in feedback-driven environments, and as unsupervised, self-supervised, and intrinsically motivated reinforcement learning research progresses, we should soon be seeing the same results without feedback. However, rather than suggesting that the integration of self-optimizing learned optimizers, MARL environments, and reward-free training paradigms are guaranteed to converge at positively 'human-level' artificial intelligence, I would only estimate that they will exhibit *more* complex, diverse, open-ended behaviors than previous architectures.

Those are just three examples where pivotal components of 'human-level' intelligence are accentuated as if they defined it. Even the brain as a whole is often exalted too high above its underlying physiological world interface and environmental interaction experience. These other two players guide the development of the brain and the intelligence it expresses. Reciprocally, the brain acts as a forcing function to maintain order over its body and interact in its environment. The body and environment play key roles in grounding the brain's internal oscillations into metaphors of cognition, and without all three, there is no human-level intelligence. Individuals raised in stimulation-poor environments or who have underlying physiological limitations, show statistically diminished potential in cultivating intelligence.

There are many other examples where an average AI researcher's prior on "human-level" artificial intelligence diverges from the real deal, and I hypothesize those seemingly sparse cases are actually uncountable. As with physics's models, the brain has been subjected to numerous comparisons over the ages including the oscillator, the clock, the steam engine, the formal proof machine, the computer, and even the neural network. However neuroscience continually

reminds us that the Brain is something else, and while we AI researchers may surpass it in various complexity, accuracy, and recall metrics, we're still not capturing its 'human-level' intelligence.

These thoughts are not new, and even researchers in the field of human level artificial intelligence acknowledge them. They may justify their use of the term 'human-level' as a means to communicate their objective to non-specialists - including committee boards, funding agencies, and executives. However, I argue that this is where the term may be abused in its worst. We scientists recognize the nuances and history underlying our terminology, so we are often able to afford the use of ambiguous terms like 'thinking', 'attention', and 'perception' to describe the activity occurring in the brain or artificial neural network. I probably have a good idea of what you're reaching for when you say "I'm developing a human-level AI system". However, when the term is taken out of context, a person may be introduced to human-level AI with the idea that it should do *everything* a real human intelligence can. Of course the realistic engineer expects to find flaws in his artificial system and eagerly looks for them, but when a funding agency, review board, or the general public are surprised by those same discrepancies, they are not amused. Funding for everybody gets cut; public interest declines; and the anticipation built up for 'human-level' artificial intelligence has been abused. If you want to help AI continue to evolve and avoid a third AI winter, don't use the buzzword "human-level" to describe your artificial intelligence.

I see two paradigms driving the advancement of artificial intelligence. The first uses natural language and philosophy to observe and reason on intelligence. It says, "Intelligence involves a collection of discrete processes like attention, perception, memory, etc. Let me axiom-atize them into components and algorithms and watch intelligence to emerge." The latter paradigm acknowledges that biological systems consist of a heterogeneous set of mechanisms to achieve their goals, but it cannot express itself in natural language. Instead it uses formal descriptions, statistical tools, and algorithms to describe intelligence. Both involve programming and experimentation. However the flow of information from observation to next iteration parameters must bounce around through a noisy natural language channel in the first paradigm, while seamlessly optimizing via pen and computer in the second paradigm. The latter approach can be intimidating for its abstract and endless complexity. However the former demands even more caution for it borders on a 'cargo-cult' style of intelligence engineering. By that, I mean:

> In the South Seas there is a Cargo Cult of people. During the war they saw airplanes land with lots of good materials, and they want the same thing to happen now. So they've arranged to make things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the

> way it looked before. But it doesn't work. No airplanes land. So I call these things Cargo Cult Science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land.[3]

Obviously, this would be a blunder when aiming to extract the principles underpinning human intelligence and engineer them into a artificial system. Admittedly, both paradigms mentioned above represent extremes between which most machine learning research sits somewhere in the middle. However, as we develop increasingly advanced artificial systems, it becomes increasingly necessary to acquire and utilize a more mathematical oriented framework of intelligence – instead of speaking about philosophically-defined axioms of thought and components of intelligence. When the time comes to program a "thinking machine", there are off-the-shelf `perceive(observation)`, `decide(thoughts)`, or `attention` functions. On the other hand, shuttling mathematical statements and statistical reports between the computer and brain is a much more straightforward task, and we can render their aims into precisely-defined building blocks like `entropy`, `mutual_information`, `free_energy`, `perplexity`, or `criticality`.

It should be clear to AI researchers who sit closer to the former extreme that advancing the intellectual capacity of artificial systems demands acquiring a basic understanding of the mathematical and statistical tools used to represent real intelligence – and not just be content with applying one or two in his or her research – but like brain's predictive model ensemble, I encourage the active-learning agent who operates in research space to entertain as many principles of human intelligence as possible. Please consider starting with the following (listed in the order you may find easiest to grasp):

1. Action and Perception as Divergence Minimization [5]
2. Friston's Free Energy Principle (I recommend [4], but you may have already found a different paper on this topic.)
3. The Energy Homeostasis Principle [13]
4. The Critical Brain Hypothesis [6, 11, 2]
5. Buzsáki's neural syntax hypothesis [1]

Once human-level AI researchers acquire a differentiable framework to propagate their thoughts through the research community, vague terms like 'human-level', 'consciousness', and 'attention' are unnecessary and actually get in the way of progress. Instead research may be defined in the language of neuroscience, psychology, information theory, computer science, or entirely new machine learning vocabulary. I personally advocate paper titles as "Towards Autonomous Developmental Language-Acquiring Artificial Intelligence" rather than "Towards Human Level Artificial Intelligence". If you make the former your grant proposal, you will almost certainly meet your objectives. However your reward estimator diverges from mine on the latter (current) paper

title. Between you and me, these vocabulary differences are only ornamental, and I can understand you when you say "human-level". However, I fear these exterior word choices might cultivate paradigmatically different mindsets and approaches to engineering artificial systems that perform on-par with human intelligence over a wide range of goals.

References

[1] György Buzsáki. "Neural syntax: Cell Assemblies, Synapsembles, and readers". In: *Neuron* 68.3 (2010), pp. 362–385. DOI: 10.1016/j.neuron.2010.09.023.

[2] Dante R. Chialvo. "Emergent complex neural dynamics". In: *Nature Physics* 6.10 (Oct. 2010), pp. 744–750. ISSN: 1745-2481. DOI: 10.1038/nphys1803. URL: http://dx.doi.org/10.1038/nphys1803.

[3] Richard P Feynman. *Cargo Cult Science*.

[4] Karl Friston. "The free-energy principle: A rough guide to the brain?" In: *Trends in Cognitive Sciences* 13.7 (2009), pp. 293–301. DOI: 10.1016/j.tics.2009.04.005.

[5] Danijar Hafner et al. "Action and Perception as Divergence Minimization". In: *CoRR* abs/2009.01791 (2020). arXiv: 2009.01791. URL: https://arxiv.org/abs/2009.01791.

[6] Janina Hesse and Thilo Gross. "Self-organized criticality as a fundamental property of neural systems". In: *Frontiers in Systems Neuroscience* 8 (2014), p. 166. ISSN: 1662-5137. DOI: 10.3389/fnsys.2014.00166. URL: https://www.frontiersin.org/article/10.3389/fnsys.2014.00166.

[7] Luke Metz et al. *Tasks, stability, architecture, and compute: Training more effective learned optimizers, and using them to train themselves*. 2020. arXiv: 2009.11243 [cs.LG].

[8] Luke Metz et al. *Training Learned Optimizers with Randomly Initialized Learned Optimizers*. 2021. arXiv: 2101.07367 [cs.LG].

[9] Nils J. Nilsson. "Human-Level Artificial Intelligence? Be Serious!" In: *AI Magazine* 26.4 (Dec. 2005), p. 68. DOI: 10.1609/aimag.v26i4.1850. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1850.

[10] Deokgun Park. *A Definition and a Test for Human-Level Artificial Intelligence*. 2021. arXiv: 2011.09410 [cs.AI].

[11] Woodrow L. Shew et al. "Neuronal Avalanches Imply Maximum Dynamic Range in Cortical Networks at Criticality". In: *Journal of Neuroscience* 29.49 (2009), pp. 15595–15600. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3864-09.2009. eprint: https://www.jneurosci.org/content/29/49/15595.full.pdf. URL: https://www.jneurosci.org/content/29/49/15595.

[12] A. M. TURING. "I.—computing machinery and intelligence". In: *Mind* LIX.236 (1950), pp. 433–460. DOI: 10.1093/mind/lix.236.433.

[13] Rodrigo C. Vergara et al. "The energy homeostasis principle: Neuronal energy regulation drives local network dynamics generating behavior". In: *Frontiers in Computational Neuroscience* 13 (2019). DOI: 10.3389/fncom.2019.00049.