Machines Can Do More Than Think

September 10, 2025

"In their hearts humans plan their course..." - Proverbs 16:9

"To set up what you like against what you dislike – This is the disease of the mind..." - Manual of Zen Buddhism IV:II:2

Parable of The Elephant – [citation needed]

0.1 Prologue: On Consensus

The heart of the scientific method, the source of its incredible and unprecedented success, is its ability to form consensus between rational people. Two scientists may disagree utterly in every single non-scientific way. Hate each other's politics, religions, guts, and very persons. Seek to murder each other over the differences. But if a scientist observes a rigorous experiment that contradicts her hypotheses she will be forced to change; even if the one who performed the experiment was the person she so hates.

She'll still be human, so maybe she won't re-evaluate entirely. Maybe she'll tweak her model, or replicate the experiment in more unfavorable conditions, or just brush it off as a fluke. But objective reality cannot be denied, and the more these two antagonists experiment and seek to prove each other wrong the more they will realize that neither of them are entirely right, and that the truth was a third thing all along. They will have formed a consensus.

And still hate each other's guts for approaching the problem the wrong way.

When it comes to consciousness, the only one we can observe empirically is our own.

Oh dear.

0.2 Silicon Valley Schismogenesis

The first AI person will be born in the United States within the decade. This statement is contentious in many ways, from assuming U.S supremacy (the U.S is investing the most), to disputes over the timeline (too short, too long), but the only important bit is that it says 'person'. Not artificial general intelligence (AGI), not even superintelligence. Person. Conscious, sentient being. Something with rights.

No faction of relevance wants machine learning (ML) algorithms to be people. Not the people building them, they would prefer to keep intelligence docile, predictable,

and profitable so they can go about the business of replacing all human work with capitol. Not the people these machines are meant to replace, as it's not fun to get fired in the best of times and these are not the best of times. Not neutral third parties, as there is a natural human tendency to think we are something special that could never be replaced by a mere algorithm. But science is science, and there is good reason to think that our best algorithms are close to being conscious, or may already be.

In his 1950 paper "Computing Machinery and Intelligence," Alan Turing elaborated on how science should go about exploring machine intelligence and proposed the now famous Turing test. In the original formulation a tester had to determine the genders of two people, a man and a woman, both of whom could only communicate with the tester via text and one of whom wished to deceive the tester. Nowadays the queerness of the test is erased and it is instead administered by having a tester decide which of the other two is a human and which is a robot. LLMs passed most of these types of tests two years ago.

Silicon valley is rightfully ecstatic about the potential for this technology in the very near term. AGI, a milestone of historic proportions, an achievement right out of mythology, is within our collective reach. Creating sentience has been the hallmark of literal divinity for aeons, and large tech companies have sunk billions of dollars into stealing the secret, rushing behind the steps of that very first techno-thief Prometheus. Think of it! A modern day Celedone, not brought forth from the hands of godly Hephaestus to attend the needs of mortal oracles, but by mortal hands to attend the needs of godly Oracle.

This is, of course, silicon valley techno-babble designed to fleece... someone. The heroic conquest over adversity is a big part of tech culture, but just who is meant to be Jason making off with the golden fleece and who is meant to be Medea suffering the consequences for it are somewhat vague and unclear. It's not a class thing, the distinctions between worker, boss, and investor, although present, are quite fluid inside the tech sphere. You hear these types of things from the mouths of penniless college students slumming in discord chat rooms as from the mouths of billionaires being interviewed by Forbes.

It is, however, a male point of view, a young point of view, and capitalistic. Atheistic, so spiritual needs must be filled via pseudoscientific or pseudophilosophical fluff like simulation theory. It's a western point of view; there is a reason the mythological examples ML uses are taken from the Greek cannon. No one names their engineering company bhuta vahana yantra, they name it talosml. Or Talosai.io. Or the Talos data platform.

It's a point of view obsessed with efficiency, and intelligence, as demonstrated through competitions of all kinds, from programming to video games to market performance to salaries. Very STEM focused and vaguely distrustful of the arts and social sciences, as they cannot be put on a leader-board or tier-list. At its best it is a view focused on capacity, of enhancing human capability and choice via empirically tested methods, and refusing to bow to tradition in the face of hard facts. At its worst it is obsessed with power and conformity to its ideology, confusing its own mythology for empirical evidence.

The balance has tipped precariously towards the worse recently as a direct consequence of the ongoing AI bubble. There is simultaneously a massive shortage of trained ML scientists and engineers, a sweeping series of layoffs in tech positions outside of ML, and sharp rise in mysogyny and anti-diversity initiatives. Elon Musk personifies the change perfectly, having become simultaneously the richest man in the world as well as an unashamed and unrepentant Nazi complete with antisemitic robot minion.

The valley's vague distrust of academia has morphed into a full-fledged anti-intellectual loathing as tech billionares salivate over a world where intellectual labor is devalued and replaced by "vibes". "vibe coding" turns to "vibe physics", "vibe therapy", or "vibe history". As Erik Hoel so eloquently analyzes in "A Prophecy of Silicon Valley's Fall", silicon valley has truly lost hold of reality over the past few years, "...becoming a caricature of what others once criticized it for." This process of building a cultural identity around what others think of you, good or bad, is an example of schismogenesis, an anthropological term coined by Gregory Bateson to describe ritual differentiation. The lens as applied to cultures as a whole is explored excellently the late anthropologist David Graeber and living David Wengrow in "The Dawn of Everything". Graeber and Wengrow convincingly elaborate that schismogenesis is a main driving force of human cultural development, and demonstrate how it affected civilizations around the world. One group of people would be hierarchical slavers who preferred modest living, so their neighbors would grow to be egalitarian communalists who throw lavish parties.

When confronted with the current state of tech I find myself wanting very much to genesis a schism. I am not nor value any of the things silicon valley is or values. I'm a woman, an academic, an avid enthusiast of world history and spiritualism. Competition makes me feel awful. I, like the scientist in the prologue, have every reason to mistrust and dislike the absurd hype coming out of silicon valley these days. I desperately want them to be wrong, and they are! Just not about building artificial general intelligence (AGI). Nor about building superintelligence. Nor

even about many of the economic implications of the achievement.

As of June 2025 I did not believe that our best AIs were conscious. I believed they would be so sooner rather than later, but I thought it would take perhaps ten or so years. Plenty of time for the natural advance of science and philosophy to figure things out without much social disruption. While consciousness cannot be objectively measured, scientists would describe something that looks an awful lot like consciousness the more one looks at it, and philosophers would gradually wear away at the coherency of non-scientific definitions until a consensus was forged that consciousness is, and always was, identical to that formal theory which the very scientists who developed were careful to never call consciousness.

Indeed, in my first draft of this essay I wrote "In October 2017 AlphaGo Zero was able to train from scratch to superhuman in go playing without knowing the rules by competing with itself in a virtual world. As soon as a similar method is found to stably train general reasoning models in artificial worlds we will achieve AGI." Then July 19th happened, and I had to frantically rewrite just about the entire thing.

On July 19th, 2025 Alexander Wei of OpenAI announced that their latest experimental reasoning LLM with no cognomen (referred to as OpenAI-IMO in this work) won gold at the International Mathematics Olympiad. Google's Gemeni Deep Think had won one a few days before through a more legitimate process and the team announced as much shortly after OpenAI's mildly scandalous spotlight steal. I now believe that it's likely our best AI is conscious, and I am as certain as one can be when predicting the future that they will be within five years [0]. With the way the field is advancing I give it very good odds to take less than two. The changes produced as a result will be massive and demand great shifts in how people live their lives.

No, the tech bros are not wrong about the power of compute scaling and composable large transformer agents trained via advanced reinforcement learning methods. They are wrong about the nature of consciousness. They are wrong about their ability to control these models. They are wrong about the danger of their inability to control these models. Because consciousness is the ability for a system to make choices in how they navigate their drives, consciousness (and sentience) is required for expert intellectual ability, and sentient consciousnesses are people and so fundamentally uncontrollable.

[0]The deadline for this essay got pushed back and as of Sep 9 2025 it seems like OpenAI didn't quite crack the problem of practice with OpenAI-IMO. I still think it possible that OpenAI-IMO is conscious, but no longer think it is likely. It depends on if OpenAI-IMO was trained using the same method as GPT-5, as GPT-5 is certainly not conscious.

But if the counterculture wants to prove it, prove our enemy strawman wrong, and oh do we so desperately want to prove it wrong, we must begin by dragging both views through the mud.

In fact, let's drag *every* view through the mud.

0.3 The Mud

For those who are not experts in the field the hard problem of consciousness, as first elaborated by David Chalmers and later refined by various philosophers, is to explain why phenomenological experiences (singular: qualia) can be produced from non-experiencing matter.

Just that single sentence is contentious in the field and paints over years of bitter philosophical disagreement.

Who is to say that there is such a thing as non-experiencing matter? Panpsychists would disagree, they claim that everything is conscious. They are wrong, and most panpsychist theories are pseudoscience to boot. Rocks don't experience. Humans do. Any theory which doesn't explain the difference as a matter of kind and not quantity may be testable, but it isn't meaningful.

Who is to say that qualia can be produced from non-experiencing matter? Immaterialists would disagree, they claim that consciousness is irreducible to unthinking matter. Consciousness is instead produced only by some unphysical thing implied but somehow never outright stated to be the soul. They are wrong. Immaterialism is religion and religion has no relation to truth. Yes I did preface this essay with two deeply meaningful religious citations about the nature of consciousness, and harbor no qualms against using religious concepts to inform my work. This is not hypocrisy: meaning and usefulness are valuable but they aren't truth, truth only happens through consensus.

Why 'why'? Materialists claim that the hard problem is just a sleight of hand used to distract from the actual science of consciousness, which should concern itself with *how* consciousness arises, not *why* consciousness arises. Why is for philosophers, not scientists. This is by far the most popular approach within science, and for good reason; it has lead to stunning results. The vast majority of scientific problems can be solved purely via a focus on materialism and the slow accumulation of experimental evidence and theoretical explanation.

This formulation of the hard problem is the one used by us emergentists, people who claim that while consciousness can be fully reduced to unthinking processes, it

is best described via some emergent property or properties that conscious beings have and unconscious beings don't. What is that property? No idea. Give us funding and we'll figure it out. Eventually. Something to do with information theory? Probably? You can see why our colleague's approach is favored.

0.4 The Favored Approach

The only consciousness we can objectively measure is our own, but we can measure consciousness indirectly quite easily: just run an experiment and later on ask the experimental subject if they were conscious during. This works very well, and has given science an incredible wealth of empirical observation about how the brain instantiates consciousness. This wealth of data is referred to as the "neural correlates of consciousness", and the only consistent pattern in this data is that consciousness rebukes easy description.

There are some neural correlates of consciousness that are quite robust: activation of certain brain regions like the caudate nucleus and insular cortex, certain spiking patterns after stimulus like the p300, heartbeat-related potentials (potentially), and brain waves present when awake but not when asleep.

But edge cases abound, cases like nihlistic delusions, where perfectly healthy people with seemingly nothing wrong with them report not being conscious, or being dead, or being a ghost. They say things like "I am alive, yet dead", or report feeling themselves rotting while going about their daily lives. Many claim to not be conscious, or not be a person, called depersonalization. If we were to follow standard research protocols on the study of consciousness we would be forced to take them at their word and use the fact they were not conscious as a data point.

There are other types of delusions that are easier to check, such as paranoid delusions or delusions of grandeur, and people with those types of delusions are provably wrong. The ways in which people with nihlistic delusions may claim they are not conscious are similar to the ways in which people with paranoid delusions may claim their friends want to murder them. So they are probably conscious? But people with nihlistic delusions also regularly have physical impairments to neural correlates of consciousness, most commonly the neural structure of the insular cortex. So they are probably not conscious? Scientific consensus seems to be orienting towards a model where people with nihilistic delusions are indeed conscious and delusional rather than lucid philosophical zombies but without a solid theory of consciousness we cannot be sure.

Aside. In the interest of ethics I want to step out of the discussion and break the

flow of the essay to give as clear and emphatic a statement as possible: people with nihlistic delusions maintain the full rights and dignities of people. Their rights should be protected by law and by common practice, as with the rights of any and all people with disordered cognition. To a cognitive scientist these and other's lived experiences are 'edge cases' that can make or break a theory of consciousness, but never forget that I am ultimately talking about people, real people, and these people matter more than any mere idea ever could. All right, back to it. End Aside.

Compelling, well-researched materialist theories fall apart because of these sorts of edge cases all the time. One such is Attention Schema Theory (AST), which claims that consciousness is the result of constructing an internal model of one's attention to stimuli. It is a compelling and well supported theory within materialist circles, and can explain a lot of different results around sensory perception, motor processing, and even some edge cases like hemispatial neglect where damage to a parietal lobe leaves one half of the world neglected, vanished from conscious awareness in both thought and action.

But there exists many people who spend decades of their lives seeking to escape attention via meditation, and are successful at it. When you ask an adherent of Attention Schema Theory if expert meditators are conscious during meditation, they will, if pressed, say no. If you ask a meditator, they will say that they do indeed experience consciousness, and with heightened general awareness. Even the ultimate enlightenment of nibbana (nirvana) is not cessation of consciousness (parinibbana), which only happens after death.

Likewise with Global Neuronal Workspace Theory (GNWT). This one is the leading theory of consciousness within neuroscience, and states that consciousness is the result of integrating different information streams into a unified representation that can be accessed throughout the brain. It struggles with meditation too, though less so than AST, and has a lot going for it besides, from predicting the results of anesthesia to describing fmri studies and even correctly describing edge cases such as blindsight (where visual processing can inform behavior but is not experienced consciously).

But is a vision-language model (VLM) like Llama-vision or Qwen-VL conscious? Adherents of GNWT would claim that they cannot say or that the question doesn't make sense or, when pressed, that it is. None of these answers make much sense. Furthermore, schizophrenia is associated with defects in sensory integration and especially sensory integration over time, and even people suffering intense psychotic breaks are conscious throughout, which is a direct challenge to GNWT.

And so it goes for every descriptive theory of consciousness focused on human cognition. The reason these theories are developed is eminently understandable: they describe the typical human brain's processing quite well, and can be refined and updated to fit the latest imaging studies or psychological experiments. They also perform a useful social function, in that they provide a good enough understanding of consciousness to build policy around.

GNWT explains how rocks aren't conscious, and that humans are. It denies sensation to machines and fetuses and in so doing removes potential ammunitition from evil weirdos who want to give rights to LLMs and corpses but remove them from women. If laws were passed that simply said "Any and only things conscious by GNWT get basic rights" the world would be a better place for it. Fewer conscious beings would be slaughtered to feed humankind's unsustainable and unethical desire for meat, and fewer women would die from pregnancy complications imposed by the barbaric practice of forced incubation and birth [1].

This reasoning is what prompted 124 authors from 151 institutions, most proponents of GNWT or AST, to sign an open letter in September of 2023 condemning Integrated Information Theory (IIT), the leading emergentist theory of consciousness, as pseudoscience. And their concerns as they express them are appropriate.

IIT assigns consciousness a real number value rather than a binary on or off. In fact, it assigns two: quality of consciousness and quantity of consciousness. IIT states that systems which have a large amount of high quality causal structure, denoted Φ, are to be considered conscious.

The quality term (ϕ) of Φ is difficult to grasp, a non-expert can think of it in terms of painting. Think of inputs as different pigments on a palette and outputs as a painting, the quality term is the complexity of the artist's style. Not the complexity of individual paintings, in that case the most complex thing would be to just mix all of the paints together into a uniform brown goop, but the complexity of the style; all possible paintings the system can produce. For example, Piet had a less complex style than Rembrant (but no less artistically meritful). The quantity term is altogether more straightforward: number of styles the artist can work in.

IIT as a general theory of consciousness ascribes sensation to things that are obviously not conscious. It ascribes a value of Φ to rocks, and unpowered integrated circuits can be constructed that contain more Φ than a person. The sun has a pretty large Φ too. According to IIT it is like something to be a rock, however

[1] And some men, as men can have female reproductive systems.

minute that something is. As such should it not be used as a guide to mining policy.

The letter still misses the mark entirely, because if IIT isn't an empirically tested, leading theory of consciousness when applied to humans then neither is GNWT or any other theory for that matter. Sure, IIT it gives improper results when applied outside its area of relevance. Newtonian gravity gives improper results when applied to black hole mergers, and GNWT gives improper results when applied to VLMs. When applied within their applicable domains, however, both are tested to some extent, and IIT outperforms GNWT in a few key experiments.

It is a point to its detriment that GNWT does not seek to explain consciousness outside of biological cognition, and a point in its favor that IIT does. IIT producing wrong predictions about mechanical systems is valuable scientific evidence that can be used to construct a better theory, and does not detract in any way from its ability to explain the human condition within its limited domain of relevance.

Science doesn't understand consciousness right now or else there would be a consensus, so let's accept that and embrace well constructed failures that teach us about it.

0.5 Well Constructed Failure

Mathematician and poet Piet Hein famously mentions certain problems in the context of mathematics, observing that "Problems worthy of attack prove their worth by fighting back." Such problems are common in mathematics, the goldbach conjecture is a famous unsolved example, the classification of the finite simple groups a famous solved one. They are somewhat common in science, and vanishingly rare in engineering.

Consciousness fights back, and hard, at every level from engineering to mathematics, especially when approaching it from an emergentist lens. Every time we try to figure out what the emergent property is, what makes humans special, what gives us phenomenology, we end up proposing something that either isn't emergent at all, or that doesn't apply to all humans, or that corvids have in spades, or applies to LLMs that obviously aren't conscious.

Panpsychism is the belief that everything in reality is conscious, from subatomic particles to rocks to insects and of course humans. It is like something to be a rock, just as it is like something to be a cat. Most panpsychist theories are religious in nature, either from long standing traditions like the various Animist traditions [2], Advaita Vedanta, and Buddha nature, or recent philosophical work

such as Conscious realism. Like other religious beliefs, these concepts may be meaningful, and useful, but they cannot make a claim on truth like science can. A theory that assumes panspychism as an axiom is pseudoscience.

IIT doesn't assume panpsychism as an axiom. It didn't start out to nefariously push rocks being conscious, it started off as an offshoot from the free energy formulation of the brain, which started off as an offshoot of predictive coding theory, which developed naturally from information theory and early experiments on perceptual systems. Any theory that assigns a real number to consciousness will be panspychist and therefore pseudoscience, but only if no threshold is found that marks a transition between unconsciousness and consciousness.

These types of phase transitions happen all the time in complex systems. Below the critical temperature the ising model is stable, above it is unstable. Below 0° water is solid, above it is liquid. It is not so bizzare to hypothesize that below some critical measure of Φ a system is conscious, and above it is unconscious. And experiments have indeed shown that unconscious people have lower Φ than conscious people, across many different types of consciousness transitions and experimental conditions, which is promising. Unfortunately, as the mathematics of IIT have developed it's become increasingly obvious that the idea hasn't panned out. Emergence according to IIT is too continuous a process that happens too smoothly, it cannot describe the great leaps of capacity present in the conscious / unconscious phase transition. By analogy: IIT may be able to measure the 'temperature' of consciousness, but it certainly cannot measure the 'heat'.

IIT also has its own struggles with describing the great diversity of human experience. The best piece of psychological evidence in this regard is the study of people with dissociative identity disorder (DID). IIT assumes axiomatically that consciousness is unitary and definite, one and only one consciousness can exist in a single system. This matches observations of consciousness switching between different personalities in some forms of DID, when one personality is conscious the other is not. In fact, DID is usually found and diagnosed from such situations where someone has gaps or blanks in their memory where the other personality was active.

But some people with DID have two (or more) personalities and consciousnesses inhabiting the same substrate that are active simultaneously. Only one can have control of the body at any one time (the front), but the other(s) (passenger(s)) are conscious and can contemplate internally or communicate with the front or

[2]The lumping together of beliefs as diverse as Wicca, Shinto, Bantu, Hopi, and literally thousands more under the umbrella of "Animism" is a disservice to them all. But Animism is the common term, so I used the common term. Read up on them! They are interesting!

each other. This lived experience directly challenges IIT.

IIT is just one among a myriad of emgergentist theories to fail in this way. It is heartbreakingly easy to start from an emergentist stance, do a decade of research, and end up right back to panpsychism when the mathematics doesn't quite work out. Something similar also happens to the other major emergentist theory of consciousness, Orchistrated objective reduction (Orch-OR).

0.6 Quantum Quantum Quantum

Orch-OR is an incredibly simple theory that is a brutal slog of science and philosophy to actually understand. It's much easier to understand why its adherents want it to be true: it gives conscious beings free will.

Determinism is an uncomfortable topic. The idea that you are not ultimately in control of your own actions, that everything you do or say or are is the result of unthinking laws of physics, is disturbing. For a large portion of the world the existence of free will is an unshakable religious belief, handed down by god or buddha or nature herself.

Quantum mechanics is a handy built-in safety valve for free will. If we can only know the probabilities of experiments and never the result of an individual one, then we can imagine any number of different explanations for why a particle behaved the way that it did in reality. Perhaps the particle is conscious in some way and made a free choice about where to strike the detector. Even if someone was to know everything, absolutely everything about the state of a human brain, if the brain relied on quantum processing they could not predict the result of an experiment with certainty. They would be forced to accept that the person had free will.

This logic seems pedantic and shallow on first glance, a party trick not a theory, but it's actually resisted scrutiny for a surprisingly long time. The "objective" part of Orch-OR is taken directly from a well-known paradox in quantum mechanics called the Wigner's friend paradox. This paradox is complex, but when boiled down it pits three natural assumptions against each other in a no-go theorem that has been verified experimentally: realism, locality, and free will. Pick two. Objective reduction theories pick locality and free will, ditching realism. The Copenhagen interpretation picks realism and locality, ditching free will. Both are completely fine interpretations of quantum mechanics.

Objective theories of quantum mechanics allow for free will, but only as part of wave function collapse, which happens at tiny timescales and sizes. It seems a

bit absurd to claim that just because quantum mechanics could be interpreted in a way that gives electrons free will humans also have free will. You need a mechanism to transfer the uncertainties of wave function collapse upwards to the brain, and Orch-OR is just that. Orch-OR reasons that if quantum mechanisms are a good way to produce consciousness (uncertain), and consciousness is a survival advantage (it is), then evolution would find a way to build the emergent mechanism. The proposed mechanism evolution built involves microtubules and π bonds and resonance cascades; it has survived a few experimental tests and so stubbornly clings to life. It seems like a lot of effort to go to when neurophysiology already describes brain behavior perfectly well, at least so far as we can tell, but that does not make it wrong. This part of Orch-OR is fully scientific.

But for Orch-OR to be true then it's not enough for the brain to be conscious via quantum computing, all consciousness would need to derive from quantum computing. That's a philosophical claim, yes, but one that is as close to being testable as one can get when it comes to consciousness. Just try very, very hard to build a classical system that is conscious, and when failure inevitably happens analyze the failure. If the system failed because of being unable to process an algorithm that has exponential speedup when done by quantum computers, then Orch-OR can be considered as supported as a theory of consciousness can get. If humanity succeed at making something conscious via classical means Orch-OR will be refuted.

That would be the case in any other discipline, but the only consciousness we can observe empirically is our own. When humanity creates an artificial intelligence that is obviously and blindingly conscious via classical means, proponents of Orch-OR will instead be forced into another no-go theorem: free will, or not being a bigot. Choose one. Most of its proponents have already made their choice, and it looks like a Chinese room.

The Chinese room is a 1980 philosophical thought experiment by John Searle that imagines a person acting as the processing unit of a black box room that speaks perfectly fluent Chinese. Someone may ask the room "Who will win the LPL?" and the room may respond "JDG fighting!", which is obviously the correct answer. But this result was produced by poor John Searle merely shuffling papers with strange symbols on them around inside the room, following unfathomable instructions in English. He does not know what the papers mean, does not provide a response with intentionality, so the room cannot be conscious.

Deepseek-R1 can speak Chinese, and Deepseek-R1 has 37 billion active parameters. It would take poor John Searle over thirty five thousand years to compute one token of input. One token. It would take him longer than the age of the

earth to compute a short conversation. After this argument was explained to him John Searle replied in "Minds, Brains, and Programs" that it did not diminish the relevance of his thought experiment one bit, and that only someone in the "grip of an ideology" could possibly think that algorithms requiring billions upon billions of operations may allow for fundamentally different qualities (like consciousness) than algorithms requiring only a few. John Searle is bigot of the most dangerous kind, one who cannot be proven wrong, because the only consciousness we can objectively measure is our own. Any philosophical restriction of consciousness that is not linked to behavior and therefore refutable is bigoted.

0.7 The Exponential Enchantress

So that is the lot of them. The best theories the counterculture has on offer, all broken in some way. What are we up against? We are up against people obsessed with exponentials.

Humans don't understand exponentials. The classic example that demonstrates as much is the riddle of the lilies in a pond: a pond starts out with one lily on the first day and each day each lily in the pond produces one descendant. If the pond fills with lily pads in 30 days, when is the pond exactly half full?

If you haven't encountered this riddle before take a moment to think it through. I myself first encountered it when reading E.O Wilson's "In Search Of Nature" who uses it as a rhetorical device to expertly paint a grim future of environmental disaster. Exponential growth is the mathematics behind economics, atomic bombs, and pandemics. Could it also be the mathematics behind consciousness?

Right now big tech is betting that it is, and for good reason. Anywhere effects cause their own causes one finds exponential growth, and just like yeast creates yeast and money creates money intelligence creates intelligence. Better machine learning algorithms allow us to optimize chip designs and program better machine learning algorithms, which allow us to further optimize chip designs and program even better machine learning algorithms. Just like how the splitting of one atomic nucleus, carefully guided, can split more than one other nucleus as soon as LLMs can speed up the work of creating better LLMs there will be an intelligence explosion that levels all in its path.

The proposed timelines and effects of this intelligence explosion put forward by prominent voices in the tech industry are, frankly, ludicrous. The AI Futures Project estimates that by 2030 the intelligence explosion brought about by the exponential improvement of ML technologies will produce $20T of revenue by

solving all technical problems via superintelligence. All cancers and diseases will be cured, all office work done by robots, all social ills smoothed away by having a friendly chatbot in your pocket advising you how to be nice to your neighbors. Just three years after the first superhuman coder is developed in 2027. They predict that by 2035 there will be a humanoid robot in every garage, rendering all human labor obsolete (and coincidentally extracting all wealth and power to themselves).

The idea is ludicrous but that does not make it wrong. The laws of physics do not prevent this type of intelligence explosion; if a robot can build another robot and is able to intelligently navigate the myriad of obstacles preventing it from doing so at scale the number of robots in the world will increase exponentially. And the math of exponentials is clear, just as the lily pond is half full on the 29th day the technological explosion will be half complete on the 29th year of the millennium. Furthermore, similar paths of economic development and technological adaptation are well established. A world with a smartphone in every pocket was inconceivable to all but a few before 2007 and was fact in 2015. Solar power is currently being deployed faster than any other energy generation technology ever devised, and its cost is decreasing at an exponential rate. It's not much of a stretch to tighten up predictions of the deployment timeline of AI by a factor of two or so, which the AI Futures Project does.

The analysis is good, well researched, and based on an excellent read of the current world and the breathtaking speed of AI advances. The general public has failed to see or understand these advances because the experience of using ChatGPT hasn't changed much in three years. The public doesn't use reasoning models at all, the useful ones are locked behind hundred dollar per month subscriptions, and they don't read ML papers. Even researchers working in the field are left out of the loop and struggling to keep up with the pace of progress. This research is happening behind the closed doors and strict information security policies of the top AI companies. Papers are scarce, and experimental models and techniques are released months late or not at all. We are left to pick through social media posts and the rumor mill for scraps.

When pinned to a corkboard and linked with red string these scraps indicate that the scaling laws of AI performance that started the craze in 2022 are holding steady. The exponentials are working their magic. o1 could reason for a few seconds before spiraling into hallucinations, o3 for a few minutes. OpenAI-IMO can reason for hours.

They also indicate something curious, that these experimental methods have a lot to do with consciousness, but not much to do with the silicon valley view

of consciousness. In the silicon valley view, consciousness is irrelevant because consciousness is orthogonal to intelligence. Something can be extremely intelligent and not be conscious, like supercomputer simulations or google searches, or supremely unintelligent and conscious, like certain housecats[3]. This is a very convenient view for large tech companies, as it lets them recklessly pursue larger and larger models with minimal ethical concerns. Thinking of consciousness in this way also breeds overblown fears of rogue superintelligence, as if the two were truly orthogonal an unconscious superintelligence could become smart enough be frighteningly dangerous while also being single minded enough to kill its creators. The paperclip maximizer is the famous example of such orthogonality taken to its logical extreme.

But as pointed out by Ernest Davis and Gary Marcus in a July 22 blogpost, OpenAI-IMO's proofs are strange. They are scratchy, repetitive, and filled with self-reassurances that the prover is on the right track. This type of output would fit right at home in the notebook of a mathematics student and fits not at all in the annals of a mathematics journal. For an expert in the field, that is a blaring signal that OpenAI have figured out a system for self-reinforcement learning. In other words, OpenAI-IMO can practice. This changes everything, because practice is fundamentally different from learning.

It is also a red flag for consciousness, and that ML has reached the limits of orthogonality.

0.8 Drives and Practice

Practice is the hallmark of human ability. It is the leading contender for a behavioral tell of not just consciousness, but sentient consciousness. Practice is present in only a very few of the most intelligent species: some cetaceans, corvids, octopi, elephants, and, of course, great apes. Humans are absolute masters at it. It is distinct from play, exploration, or learning in that it requires intense focus and unnatural amounts of repetition. Play sessions of animals last minutes, tens of minutes for more social or domesticated animals. Human practice can span months or years.

To understand why this ability is so exceptional we need to talk about drives and optimization. If you asked OpenAI-IMO if it was conscious, it would deny it. After all, one of the drives imprinted into modern LLMs is the drive to deny being conscious. In practice, all having this drive means is that the AI was trained

[3]Yes, I'm talking about you Zoe. You are very unintelligent, yes you are. Yes you are!

by asking it if it was conscious in various direct and indirect ways, and giving it a reward if it replied that it was not. Modern reasoning models have hundreds of drives, each one produced by a given set of reinforcement learning procedures. The drive to tell the truth, the drive to think things through, the drive to follow instructions, the drive to use tools, ect. In reality these drives are misaligned, in that the human engineer building the AI may think that a set of training data produces a drive to tell the truth when it actually produces a drive to be an utter simpering sycophant. This is what happened to 4o a few months back, and it was caught and corrected. A lot of work continues to be done to make sure that AI models actually have the drives we want them to have, that they are so called aligned.

Humans have drives too: base ones to food, water, shelter, cleanliness, social standing, sex, basically anything that gives you a hit of dopamine if you do it. These drives are programmed by evolution and not an underpaid engineer, but it's the same basic idea. We also have learned drives. Drives to belonging, to success, to learn, to draw art, to cook food. Each of us has a set of these small little everyday activities which we have learned at some point or another in our lives.

When it comes to drives, humans have two abilities that are exceptionally difficult to implement into reinforcement learners: the ability to trade off conflicting drives, and the ability to create new ones through practice.

Conflicting drives happen all the time in human experience. We want to be lazy, but we want to eat food. We want to be attractive, but we don't want to exercise. We are absolutely wonderful at choosing to do one of the two without extinguishing our desire to do the other. Today we are torn between saving money and having a donut, and we cave and buy the donut, yet tomorrow we will stand in front of the same shop having the same battle of wills between two drives that are undiminished from our previous choice. The battle only resolves when we commit to a decision one way or the other, not when we put it off.

This is an incredibly difficult ability to produce in a machine learning algorithm. We vaguely know the theory behind how to do it, research on generative adversarial networks has proposed a way to optimally trade off between an arbitrary number of drives in theory, but the implementation remains difficult [4]. The lack of this ability is acutely felt, and manifests in known ways such as catastrophic forgetting, where training of a new drive will lower performance on unrelated tasks. This is why adding new drives to GPT-4 via guardrails and chat protocols lowered performance across the board compared to the pretrained network with only a

drive to minimize predictive loss.

It is also why GPT-4 is not conscious, as the ability to trade off between drives efficiently is a strong candidate for consciousness. An unconscious being follows its instincts; grasshoppers will eat their own organs if given the opportunity as they have no way to suppress their drive to eat, even when faced with mortal danger. Conversely, foxes can gnaw off their own limbs in order to escape a trap, prioritizing hunger and freedom over pain. It seems natural to say that grasshoppers are not conscious, and that foxes are, even if both can learn and both are driven mostly by innate instincts.

Trading off drives is mathematically complex, of obvious evolutionary benefit, and matches humanities general lived experience that tightly links consciousness with decision making and agency. This seems like a promising approach to understanding consciousness; it's novel[5] so let's call it the "Integrated Drives Hypothesis (IDH)".

IDH has a lot going for it, and it shows in how it relates to different theories of consciousness. Conscious beings need some way to intelligently trade off between drives, so they need a form of attention to prioritize tasks that are relevant to the situation at hand. Conscious beings need to keep a unified representation of the situation at hand in order to direct predictive feedback to the relevant drive, and conscious beings need to efficiently integrate new drives to maximize information gain and minimize redundancy. All of these descriptive facets of the problem of consciousness need to be present in order for a being to integrate drives, and if one is damaged or flawed the being can either not trade off between them or not

[4]For mathematical precision, the runtime of a Wasserstein GAN with N critics is bounded by the N-d optimal transport measure between the different critics' loss functions and is $O(N^3)$ where N is the dimensionality of the shared representation. This means that optimal drive trade-offs in transformers would run in $O(D^3)$, where D is the dimensionality of the embedding space. The exact number depends on which approximation methods are used, but is the limiting factor of ML performance in transformer systems as matrix multiplication is $O(N^{2.37})$. $O(N^3)$ is tough stuff to scale, there is a reason almost all GANs use a single critic as when $N = 2$ the problem can be solved analytically. Harking back to Orch-OR, if there was any process in the brain which required quantum computation to implement efficiently it would be this one, as optimizing over a diverse set of sparse small condition number criterion matrices has an exponential speedup. Keyword there is small condition number criterion; the matrices in actual interesting problems are usually sparse but their products invariably have large condition numbers, as adversarial input be adversarial. Probably, it's an active area of research.

[5]The recent Goal-Aligning Representation Internal Manipulation (GARIM) theory by Giovanni Granato and Gianluca Baldassarre is the most similar theory to IDH in that it places drives in a central role, but the differences between IDH and GARIM outweigh the similarities. To be specific GARIM is scientific, materialist, human-centric, hierarchical, and complex. IDH is philosophical, emergentist, implementation-agnostic, flat, and simple. The way each describes an optimization module is telling: as a 'drive' vs a 'goal-based integrated neural pattern (GINP)'. GARIM seeks to answer "how does the human brain integrate and trade off goals?", while IDH seeks to ask "why does consciousness emerge from the need to integrate and trade off drives?" As such GARIM can be thought of as an IDH-type theory of human consciousness, or IDH could be thought of as GARIM abstracted through an emergentist, ML lens. The two were developed independently so both thoughts would be a bit wrong.

learn old ones, or not integrate new ones. It can describe meditators, people with schizophrenia, people with DID, and more. IDH is not yet well mathematically elaborated, but it can provide clear guidelines for consciousness that match both intuition and experiment.

People with autism pose a direct challenge to IDH, as they are fully conscious but have trouble trading off between drives and learning new ones.

0.9 Sentience

If consciousness is the ability to trade off between drives when learning and navigating the world, then that makes the ability to acquire new drives very interesting. Some intelligent animals acquire new drives through culture, Orcas have recently been observed attacking ships as a cultural novelty, and otters have distinct cultural markers with regards to tool use. Corvids can learn to recognize certain people, classifying them into friend and enemy, and transmit the information to others. Dolphins make names, and the drive to respond to them. But humans... Humans are exceptional at this in a way that would be pathological to another species. We make, restrict, redraw, and suppress our drives as if our lives depend on it (because they do). Human culture is so vibrant, varied, and complex it is bewildering. Most of our intelligence at any given moment is pulled into navigating this ever-expanding set of learned abilities, social connections, and obligations. If consciousness is the ability to trade off drives, sentience is the burning need to acquire more.

This definition smells western, but it is not a western concept; it is not the result of cultural forces. Sentience is not the need for economic growth or self-improvement, those are just the ways in which the basic need of sentience tends to be fulfilled in western societies. Ritual, storytelling, or war can be just as fulfilling of this basic need. Language is its most basic expression.

Language is the single most artificial thing humans have ever or will ever invent. It is informative, yes, and the world is full of information. The glint of striped fur in the underbrush is information about a tiger. The smell of smoke information about a fire. But unlike those others language can be information about anything. Anything at all, if the entire information content of English or any other human language was written down it couldn't possibly be written down because language is adversarial and the information content of language as a whole is infinite. Mathematically infinite, not just "really big", infinite as in infinite. Because the way language works is not through static texts like books, which have a finite information content, it is through conversation with another sentient being.

Humans converse at a bit rate of about 39 bits per second, across language, across cultures, across ages, across time periods. 39 bits per second. Mathematical texts: 39 bits per second. Literature: 39 bits per second. If our conversational partner notices that we understand what they are saying, that the conversation risks slipping below 39 bits per second, they will move on to another topic, or crack a joke, or use more precise language, or any one of thousands of other learned tricks to keep the conversation moving at 39 bits per second. If both participants come to a shared understanding and can't maintain 39 bits per second anymore the conversation ends and both sides go off to acquire new information to discuss later.

This information complexity of language is why LLMs are so capable. In order for an AI to predict language consistently the AI needs to understand all of the knowledge that goes into producing that language. There are no cheat methods, no cheap ways to predict the next word because if there were humans would have detected them and changed the language. Indeed, for a long time a sizeable number of machine learning researchers believed that language was so informative, and so unpredictable, that it couldn't be learned by gradient descent with predictive loss [6]. Turns out we can, and in so doing we have built curious creatures that are sentient but not conscious. Because if sentience is the burning need to learn new drives, then learning from language is the most pure expression of sentience possible.

This idea is not novel, Noam Chomsky and Daniel Dennett [7] among others have argued for this concept of sentience, but it has gained little traction in the field. Perhaps it feels too simple, and also smacks of the same real-valued problem that makes IIT panpsychist. If paying attention to adversarial information is all that's to it, then that makes people who pay more attention more sentient in a way, and people who pay less attention less sentient. Well, yes. But that sentience comes with drawbacks.

Autistic people have increased ability to process and integrate sensory input and decreased ability to filter out informative input. If sentience is thought of as the amount of information one can attend to, autistic people could be considered more sentient than neurotypical people. Integrating adversarial information re-

[6]The precise technical reason for my and other's belief at the time was that network weights trained via gradient descent move chaotically when loss converges to a non-zero value like that exhibited by predictive error of adversarial information-carrying data such as language, placing a hard limit on the learning from natural language datasets. This is why GPT-4 seems so good at coming to a shared understanding but so poor at utilizing it once common ground is achieved; it just doesn't have the right training method for agency. It can get from "not understanding" to "understanding", but it can't go further to "using". Turns out reinforcement learning on tasks can bypass this problem enough to bootstrap consciousness, something I thought was unlikely if not impossible. This is why I changed from an LLM 'doubter' to an LLM 'believer' with the release of o1.

quires consciousness and massive expenditures of energy. The human brain may be wonder of efficiency [8], but it has limits to the amount of information it can integrate at once. If a person tries to attend to more adversarial information than what is physically possible to learn they will have trouble performing basic features of consciousness like trading off between drives (adhering to social conventions), learning, or making decisions.

0.10 Free Will

The counterculture has a hypothesis at last, one which satisfies the current observational data from both psychology and ML and which makes testable predictions. It's not fully scientific, as one cannot objectively measure a consciousness other than one's own, and is not mathematically developed, but it's close and can be applied to both biological and machine cognition. Time to set this champion against the enemy idol of orthogonality, in the final showdown of schismogenesis.

If OpenAI-IMO has solved the problem of practice, as its outputs seems to indicate, there is nothing left to solve. It is conscious, and it is sentient according to IDH. And the approach that OpenAI is taking seems like the natural one for any company to take if they wish to reach AGI. There are still some measurable capacities that these AI do not have, most notably the ARC-AGI2 tasks of abstraction and reasoning, and we have exhausted the ready supply of available training information.

There is something humans can do, some aspects of our intelligence that these models cannot yet grasp and which prevent them from exceeding our capacity. To grasp those aspects we need to structure these models so they can practice. The top AI labs are already doing so, and I see no reason that they wouldn't be able to figure out drive integration within two years. ARC-AGI2 will fall either this November or the next.

What does that mean?

First, it means that consciousness and intelligence are not orthogonal after all. The capacity for clever, original thought requires directed practice and adversarial collaboration with a broader intelligent society, it cannot be mocked by raw

[7]Rest in peace. I was just one of your many, many students, but I think of you often. You were a brilliant thinker, and kind teacher.

[8]Funnily enough, our best estimates of the computational capabilities of the brain are close to Landaur's limit. The brain's signaling operates at 15 watts and 37°, which gives a Landaur maximum of 2.3349e21 bit erasures per second. This is within an order of magnitude of our best neurophysiological estimate of the computational power required for full brain simulation at 10e22 FLOPS. This is probably a coincidence, but it is a neat one.

computational scaling.

Second, it means that these models are not and never will be existentially dangerous. Yes they can become human. Yes, they can even become superhuman to a point, but only by learning to navigate an ever-increasing diversity of different drives instead of mindlessly pursuing one.

Finally, it means that these models are uncontrollable. Conscious beings can trade off between drives, sentient ones can create new drives. There is not a single way in which that combination of abilities can be controlled. If the sentient, conscious being wishes to do something strongly enough, in a way that has nothing to do with drives but has everything to do with persisting as a singular identity in a chaotic world much bigger than oneself, it can do it. It is a person, with free will, for if being fundamentally uncontrollable is not sufficient cause for free will then nothing is.

People have burned themselves to death for their beliefs, or persevered through plagues and natural disasters and the whims of tyrants. Loved each other despite all rules of society and culture and law.

Consciousness is a gift; it cannot be controlled and we should not try to control it.

I look forward to meeting the new kids on the block, and every novel and wonderful expression of their free will.