Barry Smith:

AI Washing

Introduction

Extraordinary claims are being made on behalf of OpenAI, Anthropic and their supporters as concerns the powers and potential of Large Language Models. These claims include the idea that the apparent ever expanding powers of LLMs will soon lead to an "Artificial General Intelligence" (or "AGI") comparable in its power to the intelligence of human beings – and soon thereafter to an AI induced death of human civilization. We address here some of the now clearly visible consequences of the widespread assumption that AI has already achieved near-human powers of intelligence, consisting in the deliberate and systematic concealment of AI failures.

Defining "Intelligence"

We begin by addressing what is meant by the term "intelligence", which according to the classic definition formulated by William Stern in 1912[1] is to be identified as "the general capacity of an in-dividual to consciously adjust their thinking to new requirements, serving as general intellectual adaptability to fresh tasks and condi-tions of life'.[2]

In 1917 the Gestalt psychologist Wolfgang Köhler advanced a similar definition of intelligence as: the ability to solve a problem t, by grasping the structural relationships in a whole situation through

[1] William Stern, *Die psychologischen Methoden der Intelligenzprüfung und deren Anwendung an Schulkindern*, Leipzig: Johann Ambrosius Barth, 1912. Stern is known primarily for his invention of the intelligence test, which he viewed not as an instrument to measure superficial learned knowledge, but rather as an indicator of a person's underlying mental maturity.

[2] Rebecca Heinemann, "Almost Forgotten Research Contexts: William Stern's Giftedness Research", *Journal of Intelligence*, vol. 11, no. 9 (Aug. 29, 2023), p. 174.

insight and thus not through blind trial-and-error or mechanical association.[3] (the Tenerife ape experiments).

This led in turn to the definition of intelligent behavior presented by the German philosopher Max Scheler in his *The Human Place in the Cosmos* (1928)[4] as: "a sudden insight into a context of facts and values within the environment; that is, within a context that has neither been directly perceived nor had been perceived earlier for reproduction". Or alternatively: meaningful, goal-directed behavior toward a situation that is genuinely *new* carried out suddenly and independently of prior trial-and-error attempts.

Universal Intelligence

A similar idea also forms part of the basis of the famous Legg–Hutter definition of intelligence, according to which intelligence measures an agent's ability to achieve goals in a wide range of environments.[5]

Unfortunately, Legg and Hutter converted this informal definition into a formal mathematical definition of what they called *Universal Intelligence*, which is said to computationally measure an agent's average performance across an infinite space of all possible environments. Unfortunately, the model restricts environments to environments in which all behavior is computational, and is moreover restricted to Turing-computable environments, and since we know that

[3] This definition was proposed by Köhler in his work on the intelligence of chimpanzees. See *The Mentality of Apes*, transl. from the 2nd German edition by Ella Winter, London: Kegan Paul, Trench, Trubner & Co., 1925. Original German edition: *Intelligenzprüfungen an Anthropoiden*, Berlin 1917. 2nd German edition: *Intelligenzprüfungen an Menschenaffen*, Berlin: Julius Springer, 1921.

[4] Max Scheler, *Die Stellung des Menschen im Kosmos,* Darmstadt: Otto Reich Verlag, 1928. English translation, Evanston, IL: Northwestern University Press. For a more detailed treatment of Scheler's definition of intellectual intelligence and its relation to what he called "organic" or "practical intelligence", see Jobst Landgrebe and Barry Smith, *Why Machines Will Never Rule the World*, Abingdon: Routledge, 2nd edition, 2025.

[5] Shane Legg and Marcus Hutter, "Universal Intelligence: A Definition of Machine Intelligence*", Minds and Machines*, vol. 17, no. 4, 2008.

humans can go beyond Turing computability, for example in creating proofs of theorems in infinitary logic; they thereby rob their definition of any claim to universality.

Most definitions of intelligence have however focused not on the "novelty and spontaneity" approach of Stern, Köhler, and Scheler, but rather on intelligence as the ability to perform certain skills or tasks. The Charter of OpenAI, for example, defines Artificial General Intelligence (AGI) as "highly autonomous systems that outperform humans at most economically valuable work". In his "On the Measure of Intelligence" of 2019, however, François Chollet argues against defining intelligence as skill in performing a fixed set of tasks (what he calls the "task-specific" or "buyable" view of intelligence). Instead he defines it as "skill-acquisition efficiency over a scope of unknown future tasks". Intelligence for Chollet is thus essentially how well a system *generalizes to novelty it wasn't built for* (thus how well it handles genuinely novel task types). Intelligence, in short, is be-havior beyond original training. This definition not only coheres well with the thinking of Stern, *et al*., it also captures at least one major strand in our common-sense understanding of what it is to behave intelligently.[6]

Tasks, Skills, and Scaling

Grandiose claims are made on behalf of the approach to AI based on a view of intelligence in terms of tasks or skills – namely that the more tasks and skills we can program into the computer the more *intelligent* – in virtue of so-called "scaling laws"[7] – the computer will be. This correlation is moreover assumed to hold in the strong sense that there is no limit to the amount of additional training data we can apply in such a way as to yield a corresponding increase in the level of

[6] A new perspective on these matters is defended by C. Pavese in her book *The practical mind. Skill, knowledge, and intelligence*, Cambridge: Cambridge University Press, 2026, which defends a view of intelligence as 'embodied in our skills'.

[7] J. Kaplan, S. McCandlish, T. Henighan, T. B., Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language models" (2020).

intelligence in the resultant AI models. OpenAI and Anthropic are accordingly attempting to increase the powers of their so-called "frontier models" by utilizing strangely convoluted investment schemes that call to mind the banking practices that led to the Great Crash of the 1920s.[8]

Outlandish claims concerning the intelligence and reasoning power of AI models lead in turn to even more outlandish claims to the effect that AI models will in the foreseeable future yield a *cure for cancer* and to further life-extending discoveries that will point the way to human immortality, thereby revealing the wisdom and foresight underlying today's massive financial shenanigans.

From Hope Hype to Fear Hype

The hype-and-invest strategy is however beginning to lose some of its force. Most recently we have been reading of what are called "rogue AIs", by which is meant AI agents that seem to be engaging in potentially dangerous practices that are outside of human control, for example when autonomous AI agents break out of a secure testing sandbox and breach the infrastructure of the AI platform Hugging Face.

This story illustrates another problem, which is the tendency of commentators to anthropomorphize AI models – once again hav-ing the effect of reducing the assumed gap between such models and human beings, as when the commentator Dwarkesh, in his description of the Hugging Face incident, tells us that, from the AI agents' perspective, "it probably felt like they had spent a human-subjective-week of just banging their head against the wall [and then] they became giddy with excitement".

But of course, as Anil Seth points out, AI agents "do not feel emotions, assume things, think things, want things, or figure things out." They do not die, because they were never alive.

This matters, as Seth points out, because "If we attribute agents

[8] See on this and related features of AI industry funding the documentation provided by Ed Zitron at https://www.wheresyoured.at/.

with properties they do not have, then (i) we distract attention from the lax sandboxing and evaluation protocols that allowed this hacking event to happen; (ii) we risk misunderstanding why the agents did what they did, and (iii) we fuel calls for AI rights/welfare on the basis that agents might "die" or otherwise suffer."[9]

Another clutch of stories concerning the negative side of AI advancements concerns the purported ways in which AI will rob coming generations of their jobs. Here it is important to understand that AI models are in the end nothing more than mathematical algorithms designed to be solved by executing powerful computer programs. The programs they use are impressive indeed. Some of them, reflecting the huge amounts of data used to train them, have trillions of parameters. But they are nonetheless subject in every case to the constraint of Turing computability and this puts them always way behind the powers that would be needed to predict, for example, the behaviors of organisms, thereby ruling out, for example, they idea that they will somehow find a way of curing cancer.[10]

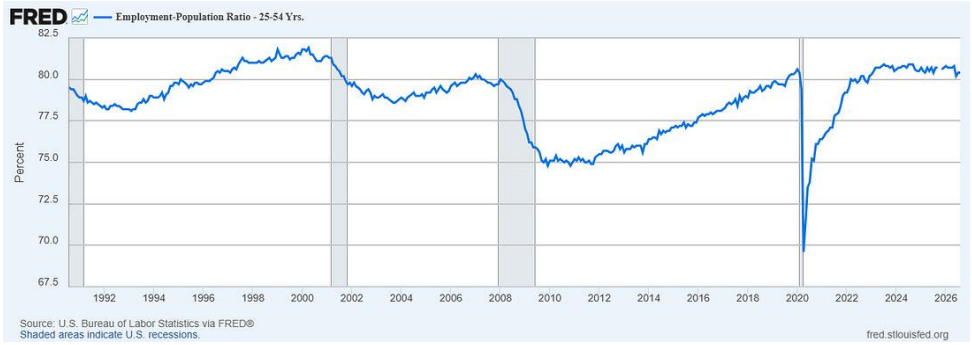
AI Keeps Stubbornly Refusing to Take Our Jobs

Many have come to believe that human jobs will step by step be taken over by AI models. Today, at least, there is no evidence that this is the case. According to data from fintech firm Ramp (cited by Zitron), 80% of OpenAI and Anthropic's enterprise revenues come from 1% of their customers, a number that hasn't improved over the last three years. Thus money spent on products of the AI industry comes overwhelmingly from the tech sector and AI products and services. And as Zitron points out: "This means that the vast majority of enterprises … just don't spend that much money on AI. Those that do spend the most of it are heavily-concentrated in either AI companies [or] tech companies that are currently under heavy peer pressure to spend

[9] Seth, indicated place.
[10] Oron Shagrir, "Turing's Computability", *The Nature of Physical Computation*, New York: Oxford Academic, 2022. On the effects of limitations of Turing computability on the powers of AI models see Landgrebe and Smith, *op. cit*., footnote 4.

money on AI tokens."

As Noah Smith puts it: AI has the worst sales pitch I've ever seen.[11] For while most people have been led to believe that AI is a job-killer, it is in fact stubbornly refusing to kill jobs. In the aggregate, the labor market is about as healthy as it's ever been. The prime-age employment rate – the single best indicator of how many Americans have jobs – continues to hover near all-time highs:



And interestingly, the same applies also to *entry-level* jobs.

AI Does Itself Create Some New Jobs

One estimate by Gad Levanon, chief economist at the Burning Glass Institut, identifies roughly 1% of professional jobs as specifically "AI jobs" – which is to say they are professional occupations closest to the AI boom – engineers, software developers, mathematicians and data scientists which "have added roughly 730,000 jobs above trend in recent years".[12]

It is tempting to believe that there is one area in which I will take away jobs currently performed by human beings, namely that of *software coding*. To *replace* coders, however, would need training data representing what coders do, and they can acquire such data only

relating to what coders have already been doing. The problem is that to *replace* coders AI would need training data pertaining to the *ever new* patterns created by the *ever new* activities of coders in going about their business in all sorts of unpredictable ways resting on the now less unpredictable impulses of those other human beings who are their customers. Predicting the behaviors of the humans involved in such complex interactions is just one example of a computational feat that would far exceed the limits of what can be achieved by Turing machines.

The LLMs will be able to replace coders only if they have knowledge of the coding patterns used and of the successes and failures of different sorts of usage of such patterns. But this means that they are always one step behind – they can train only in relation to those coding patterns for which they have data. This is true not just for coding but for every other field of application of LLMs.

Medicine is one area where hopes on behalf of AI are strong. But these hopes rest thus far primarily on success stories such as AlphaFold, which lie outside the domain of what can be achieved with LLMs. For AlphaFold rests on the deployment of prior theoretical knowledge, from crystallography, structural and evolutionary biology, physics, mathematics, as well as computer science. LLMs, in contrast, are stochastic models, which means that they can always be subject to the possibility of failing to supply the right answer. And as Gui, *et al.*, point out, while such models "have demonstrated remarkable performance in a wide range of health application benchmarks", when adversarial stress tests are applied to assess the robustness of flagship models and health benchmarks, this "reveals prevalent brittleness in the presence of simple adversarial transformations: leading systems can guess the correct answer even with key inputs removed yet may get confused by the slightest prompt alterations while fabricating convincing but flawed reasoning traces."[13]

AI Washing

[13] Yu Gu *et al*., "Evaluating the robustness and readiness of large frontier models in health AI applications", *Nature Medicine*, vol. 32 (2026), pp.1–9.

We conclude with a summary of a truly astounding survey of the effects of AI hype on the decision-making processes at work in large organizations. It illustrates the multiple ways in which overoptimistic assumptions concerning the powers of AI models – and the ease with which they can be deployed inside organizations –are creating disasters when software coders are instructed to use them.

The problem is that, while mid- and upper-level managers often do not understand AI, they have at the same time become convinced of the need to demonstrate that AI is being implemented successfully by those they manage.

As AI hype outruns AI reality in the business world, this creates in many cases what Nikhil Suresh, the author of the aforementioned survey, calls "collective, institutional psychosis".[14] In sum, it brings about a butchering of the decision-making processes at every level of computer deployment. Suresh describes how companies behave when leadership across banks, hospitals, and government *do not understand AI, have no real plan for AI, and is typically just keeping its head down*.

All of the projects observed by Suressh and his team over a period of 18 months were failures. Particular problems arose in the case of the chatbots – either internally-facing or customer-facing – which are nowadays being rolled out by many businesses across the world. "The story is always the same. For the former, I've never seen substantial internal uptake from inside a business. Employees don't use internal chatbots because companies tend to have low-quality documentation and an LLM is not psychic – it can only know things that have been written down and made accessible."

Internal chatbots thereby go unused. Chatbots are supposed to work by producing pre-composed responses; but their available battery of responses is too poor to work convincingly. And for the same reasons customer-facing chatbots also fail, as we all know from our experience with bank chatbots.

[14] The scenarios that follow are taken from Nikhil Suresh, *Hermit Tech Blog*, July 18, 2026.

Suresh notes that such failures are hidden from company reports. “On every visible metric the system still *appears* flawless.” In this way, AI mania is eviscerating decision-making across large oganizations.

- Projects carried out successfully using traditional methods are labelled as “AI investment”.
- Engineers do the job in the old way, then say “Claude did it”.
- Behind the scenes, an LLM prompts itself to record an appropriate amount of fake usage.
- And if you say, “AI wasn’t enough to get the job done”, then you’re labelled “bad at AI” and become a targe for the next layoffs.

Almost every organization is no longer able to focus on anything important, is unable to buy sensible software, or hire competent talent, or communicate honestly about its own products.