

Creation Ex Nihilo

On the possibility of autonomous progress and recursive self-improvement

Camin McCluskey

2026-08-14

This essay argues that contemporary language models, despite their extraordinary predictive and generative power, do not yet possess the capacity required for autonomous progress. Progress being the creation of new knowledge through good explanations: hard-to-vary conjectures that survive criticism and reach beyond their original observations. For this reason, current AI systems can accelerate human inquiry, automate vast regions of cognitive work, and serve as powerful collaborators, but they remain structurally limited by statistical prediction, borrowed causal perspective, and entangled representations. The central question is therefore not whether AI will become useful or transformative, but whether predictive systems can ever discover the generating process of knowledge itself.

Contents

Introduction	3
Progress: Humanity's Collective Project	4
Why is the winter cold?	5
Progress as the fractal creation of good explanations	10
Imago Dei: Instantiating Intelligence	11
Statistical Learning as a Universal Explainer	13
Entropy	14
The Distribution Hypothesis	17
The Bitter Lesson	18
Finding the Generating Process of Knowledge: Scaling the Causal Wall	19
Causal Inference	20
External Sense Makers, Simulation, and Entangled Representations	22
Simulation	23
Fractured Entangled Representations	24
Implications	25
Transient Human Advantage	27
Conclusion	29

Introduction

The mystery of the human mind is one that has captivated philosophers and scientists since antiquity. How is it that we know ourselves, and imagine others, to have consciousness? Do we have free will to decide our actions, or is that an illusion? How is it that a bundle of cells gives rise to the rich interiority of being a human?

Throughout history, humanity’s best understanding of these questions has been to draw analogies to the frontier technology of the age. Whether that was clockwork mechanisms, hydraulic pipework, mechanical calculators, or telephone exchanges¹. At each point we have been wrong, which is to be expected. As I shall explain, the enduring project of humanity—a project we call Progress—is the invalidation of bad explanations in favour of better explanations. In the spirit of this project, the explanations that prevailed in each successive era *have* been better explanations for human cognition.

However, a very critical sense, to analogise the brain as a computer is unlike any previous comparison. In contrast to all prior technology, a modern computer is a *universal simulator*—a machine which can simulate the behaviour of any physical object to any desired accuracy when provided with sufficient memory and compute. The brain is such a physical object, albeit one of enormous complexity, and we presume it must obey the laws of physics like any other object. Therefore a computer can, in principle, simulate a human brain up to and including every process—creativity, imagination, love, even consciousness—contained within.

Neural networks, and now Large Language Models (LLMs) are software abstractions which run atop this universal computing hardware. So our best explanations of how to emulate or instantiate cognitive processes are now ones of computer software abstraction. Finding the algorithms of consciousness, intelligence, and everything else is the latest chapter of our collective project.

The purpose of crafting these explanations has always been to use this understanding to create artificial brains, free from biological constraints. Depending on who you ask, this ambition, once the preserve of science fiction, has either been achieved in practice or achieved but for scale. LLMs from 2022 onwards pass the Turing Test, once seen as a definitive proof of intelligence. LLMs “know” more than the average person, and increasingly more than experts in many fields. They can aggregate information from across the web in seconds and act upon the world in increasingly sophisticated ways. They continue to exceed human performance on genuinely hard benchmarks. Benchmarks which appear to require not just knowledge, but reasoning and intelligence.

It is tempting to conclude that language models have now crossed the threshold of general intelligence. Or at least they can approximate intelligence to a degree that will soon render human cognition unnecessary for most-to-all economically valuable work. In short order we

¹Andrew P. Wickens, *A History of the Brain: From Stone Age Surgery to Modern Neuroscience* (London: Psychology Press, 2015).

should expect frontier labs will turn the capabilities of their models toward improving themselves in search of recursive self-improvement. If this works then humanity is forever cut out of the project. Our cognitive services, at the very least, no longer required.

I believe the future looks rather different. That, despite appearances, the era of human cognitive primacy will continue against a counterintuitive backdrop of AI systems which crush every benchmark and automate the vast majority of work humans do today. To believe this requires understanding what we still cannot teach machines, how our current explanations of intelligence are flawed and, most importantly, how the project of Progress actually works.

Progress: Humanity's Collective Project

How does progress happen? Why do we assume life tomorrow will be better than life today? How is it that we, as individuals and as a species, solve problems that stumped previous versions of ourselves or our ancestors?

It is important to understand the nature of progress because it is the defining force that improves our lives. More than that, sustained progress is imperative to ensure the survival of our species. The definition we shall use, with explanation and motivation to follow, is best articulated by David Deutsch's *'The Beginning of Infinity'*, which has its groundings in Popperian theory.

Progress is the objective transition from misconception to ever better misconception, achieved through the quest for "*good explanations*"—hard-to-vary accounts of reality that solve current problems, reveal "better" problems, and create knowledge with universal reach.

The creation of knowledge is essential for the continuation of the project because any transformation of matter that is not forbidden by the laws of nature is achievable provided the right knowledge is present. The earth is not an intrinsically hospitable place for mankind. It has been made so by humans using our combined knowledge. We are the designers and builders of our world, not merely its stewards and caretakers. Nature does not grant us warmth, clean water or abundant food. We rely on the ingenuity of our species for these things which are essential to our survival.

Using knowledge to enact some form of transformation is the basic method by which all organisms survive and reproduce. Your genes are packets of information, encoded in which are instructions for assembly and useful behaviour. These instructions are what keep every living being on the planet alive. The difference between humans and all other species is what kind of knowledge we can use: explanatory vs. rule-of-thumb (innate) and how we create such knowledge: conjecture and criticism vs. the random variation and selection of genes. The difference between how we accumulate and deploy knowledge explains why humans have come to dominate this planet, and why there is, in principle, no reason we couldn't come to have dominion over the rest of the universe.

All living organisms are factories for transforming a small subset of the universe's abundant resources into more of themselves. Humans are singular in that we are factories for transforming *anything* into anything else, subject only to the laws of nature. Humans are “*universal constructors*”. So to understand progress, we must understand how humans create knowledge. How is that we are uniquely capable of creating and utilising knowledge to achieve any transformation that reality allows?

For much of human history—up to the period of the Enlightenment—this was not well understood. Knowledge creation happened rarely and when it did occur it was likely accidental. Chance events, or happy accidents led single individuals to discoveries about the nature of their proximate environment. What passed for knowledge were “*bad explanations*”—parochial tales which purported to explain facts about the universe. These explanation were, at best, limited explanations of local phenomena. More commonly, they were just plain wrong. As such, they were an unreliable substance with which to build progress. This was the situation for virtually all of human history, when life changed almost imperceptibly from one generation to the next.

It was not a physical difference that held our ancestors back. Our brains and their brains are virtually identical. The difference is epistemological. They has the wrong idea about what knowledge is and how it grows.

Why is the winter cold?

Every civilisation in the Northern Hemisphere has faced a question of why the months of winter are perceivably colder than all other times of the year. Each civilisation has likewise proffered an answer to this question. It's through these answers and the subsequent questions they spark that we can reveal how they thought knowledge worked.

Long ago, Hades, god of the underworld, kidnapped and raped Persephone, goddess of spring. Then Persephone's mother, Demeter, goddess of the earth and agriculture, negotiated a contract for her daughter's release, which specified that Persephone would marry Hades and eat a magic seed that would compel her to visit him once a year thereafter. Whenever Persephone was away fulfilling this obligation, Demeter became sad and would command the world to become cold and bleak so that nothing could grow.



Figure 1: *The Rape of Proserpina* (Peter Paul Rubens, circa 1636–1638)

For the Ancient Greeks this *was* the explanation of winter. Not an analogy, or a decorative illustration of the principle, but the answer itself. It’s clear to anyone living today that there is something flawed about this explanation. But what exactly is wrong with it? It’s certainly *an* explanation for the phenomena of the coldness of winter. How can we conclude it’s false?

You might suspect it is because the explanation is not testable. But that is not the issue. The fundamental flaw in this explanation is that it is *easily variable*. This is the reason that myth-making in general is not an effective way to understand the world. This is true whether the myth is testable or not, for whenever it is easy to vary an explanation without changing its predictions, one could just as easily vary the myth such that different predictions are produced, as required. For example, if the ancient Greeks had discovered that the seasons in the northern and southern hemispheres are out of phase, they would have had a choice of countless slight variants of the myth that would be consistent with that observation.

In general, when theories are easily variable, experimental testing is almost useless for correcting errors in the theory. Testing does not help us progress towards the truth when the theory under test is a “bad explanation”.

All myths in this period of human history, regardless of detail, eventually resolved to some version of “the gods did it”. Such explanations actually explain nothing because they could be used to explain *anything*. This is more generally true of any society which believed that knowledge flows *downward* from gods, sacred texts, religious leaders, or tradition. For them, to know something was to receive it from a “legitimate” source. To achieve “justified true belief”², where the source of justification was axiomatic. Such societies are ultimately static. They have cultures in which deviation from inherited knowledge was not only wrong but dangerous. These knowledge structures are self-sealing. You cannot not criticise them without already violating them.

²Ichikawa, Jonathan Jenkins, and Matthias Steup, “The Analysis of Knowledge,” *The Stanford Encyclopedia of Philosophy* (Spring 2026 Edition), ed. Edward N. Zalta and Uri Nodelman, <https://plato.stanford.edu/entries/knowledge-analysis/>.

When secular thinkers broke from this tradition, often at great personal cost, it represented a major step forward for the growth of human knowledge. For the most part however, they did not appreciate the value of explanation but instead deferred slavishly to data. They fell into *inductivism*. This is the misconception that scientific theories are obtained by generalising or extrapolating repeated experiences.

Every year without exception, December, January and February are cold. March, April, May are warming. June, July, August are hot. The pattern repeats, year after year, across recorded history. Therefore: winter is cold because that is what winter does.

Inductivists imagine a “principle of induction”. That the *unseen resembles the seen* or that the *future will resemble the past*. This feels like progress. It is grounded in evidence. It makes testable predictions. But it is wrong. Travelling to the southern hemisphere disproves the theory instantly. The pattern, which was reliable for millennia in the northern hemisphere fails instantly. The inductive generalisation “winter months are cold” was never a true explanation at all. It was a local description mistaken for a universal law.

As a consequence, a good explanation must also have *reach*: It should account for what you have observed and make testable predictions about cases you haven’t seen, including those that contradict your prior observations. An inductivist has no framework for understanding why Cape Town is hot in December because they never had an explanation to begin with, only a pattern.

A more sophisticated thinker (and to surprising degree many AI research labs) might say “fine, we just need more data”. The issue with the explanation that winter is always cold is that simply observing the temperature doesn’t account for the full richness of reality. Given that we know the sun produces the heat on earth, let’s measure the position of the sun relative to earth during the year. Our explanation then becomes:

In winter, the sun is lower in the sky and spends fewer hours above the horizon. In summer, it is higher and stays up longer. We observe this consistently. Therefore: it is the sun’s behaviour that causes winter.

This is no longer an strictly inductivist theory but instead an generally *empiricist* view. Our explanation is derived directly from experiential data.

The flaw in this explanation will be familiar to any high school statistics student—correlation does not imply causation. The fact that the sun is lower correlates with colder temperatures does not tell us anything about *why* the sun is lower in the sky. Does the low sun cause the cold temperatures? Do they instead share a common cause? Ancient astronomers observed the sun’s arc with extraordinary precision. They were able to predict solstices and equinoxes to the day. And yet for over a thousand years, the dominant explanation was that the Earth was the centre of the universe and the sun moved around it. Comprehensive observation can still lead to flawed theory.

The trap that the empiricist falls into is forgetting that all observation is *theory-laden*. It is not possible to collect neutral facts from which explanation can be derived. It is never possible to “let the data speak for itself” without smuggling in assumptions you have not examined. What counts as relevant? What is background and what is signal? What counts as a category in our data collection?

At this point you may wonder; does “why” really matter? If we have a model which predicts that which you need it to predict to an satisfactory accuracy then what more could you want?

We don't need to know what “really” causes winter. We have equations that predict it perfectly. Science is about prediction, not explanation.

This is the *instrumentalist* view. It's obviously seductive, particularly when our models become very accurate. Perhaps it is what science *ought* to be. More calculation, less storytelling. But the issue with instrumentalism is that gives up on understanding, and therefore progress. If we become content with predictions alone, even as they increase in accuracy, there is no reason to look for better theories. The instrumentalist has no criterion for progress except accuracy of prediction. By that standard, the wrong model can be “good science” indefinitely. Critically, the instrumentalist approach blocks the next question. When the model “works” we stop asking what the model is actually tracking. This is bad science and it must be rejected.

The thinkers of the Enlightenment rejected bad science and in doing so unleashed the project of Progress by formalising several philosophical shifts:

1. **Rebellion against authority:** The Enlightenment was, at its core, a rebellion against the idea that knowledge comes from “legitimate” source like ancient texts or religious leaders. Indeed the motto of the Royal Society is “*Nullius in verba*” (“Take no one's word for it”).
2. **Leaps of conjecture:** Perhaps counterintuitively, science could only progress when it moved from fitting data to boldly *guessing* a new reality. Scientific theories are not “derived” from anything. They are conjectures.
3. **Tradition of criticism:** Rejecting authority alone is not enough. One must cultivate a tradition of criticism to systematically detect and eliminate error in conjectures.
4. **Requirement of Testability:** Testability arises not as a source of truth, but as a specialised form of criticism used to choose between competing “good explanations” that are already hard to vary.

It is through these traditions that we can understand a good explanation for winter.

The Earth orbits the Sun while tilted on its axis at approximately 23.5 degrees relative to its orbital plane. As the Earth moves around the Sun across the year, this fixed tilt means that each hemisphere alternately points toward and away from the Sun. When your hemisphere points toward the Sun, sunlight strikes at a steeper angle resulting in more energy per unit area and longer days. That is summer.

When your hemisphere points away, sunlight strikes obliquely, spreading across a larger area, and days are shorter. That is winter.

Every detail is now load-bearing. Change the axial tilt and the season change in precisely calculable ways. Less tilt and you get milder seasons. No tilt means no seasons at all. Change the orbital distance and you change the baseline temperature but not the seasons. In other words the explanation is *hard to vary*. Changing a single detail results in predictions that no longer match reality.

Moreover, the *reach* of this explanation is extraordinary:

- It immediately explains why the southern hemisphere's seasons are reversed without additional assumptions
- It explains why equatorial regions have weak seasons and polar regions have extreme ones
- It predicts what seasons would look like on Mars (which has a similar axial tilt but a much more elliptical orbit) before anyone had ever looked, or *could* ever look.

Throughout space and time, the explanation reaches far beyond what was seemingly contained within the original observations. This reach is evidence that the explanation has captured something real about the structure of the reality.

It's worth noting an important fact about the Axial Tilt theory. It was the ancient Greeks, in the 4th Century BCE, who measured the tilt of the earth and explicitly linked this to the changing seasons. Long before the Enlightenment, there were those that sought good explanations. But the traditions of the Enlightenment, which give rise to the political, cultural, moral and economic establishment we describe as "The West", operationalise the search for good explanations at a societal scale. We should take this as proof that progress is not inevitable. History contains numerous examples of societies in which anti-rational memes and traditions prevailed. From Renaissance Florence to Plato's Academy in Golden Age Athens, the process of knowledge creation is ever vulnerable to static societies' resistance to change, to dogma, and taboo.

Static societies inevitably result in catastrophe. Only progress is sustainable in the long term. A society that stops creating new knowledge loses the ability to recover from unforeseen failures and will eventually be destroyed by a problem it did not anticipate.

We must therefore continue the march of Progress. But perhaps it's the case that only a select few in society create new knowledge and it is the role of the rest of us to disseminate and apply that knowledge? You might argue that the common definition of "progress" could reasonably include taking what works somewhere—some scientific discovery or innovation—and bringing it into a new domain or geography. I do not believe this is progress, nor do I think the responsibility to improve the condition of our species falls to only a select few. It is incumbent on us all, and we rise to that challenge far more frequently than you might suspect. Put simply, progress is fractal.

Progress as the fractal creation of good explanations

Dawkins describes how an idea spreads in his 1976 book ‘*The Selfish Gene*’ as analogous to how genes propagate themselves in a gene pool by moving from body to body. Ideas (which he termed “memes”) propagate themselves in a “meme pool” by leaping from brain to brain via imperfect imitation. Importantly, memetic replication is a creative activity in itself. You cannot download a meme into your brain (yet at least). To understand an idea you must *conjecture* the meaning of a behaviour or utterance and then reconstruct that idea in your own mind. Anyone who struggled in maths class will intuitively understand the frustration that arises because understanding is self-constructed. You cannot learn an idea if you are missing the necessary prerequisite to understand it, any more than you can understand a joke in a language you do not speak.

For this reason we should not suspect that there are some “blessed” brains that are capable of de-novo creativity and some that are not. Every human is a *universal explainer*. Each of us has the same underlying *capacity* for creative thought and knowledge creation. There is no qualitative difference in the capacity of genius and that of an average person. Any apparent limits are those of current ability, motivation, or the suppression of creativity by anti-rational memes, including those that are self-imposed (“I am bad at maths” or “girls/boys don’t do X”).

Just as we should not reserve special importance for the brains of the few, we should not reserve special importance for fundamental science as the only *domain* in which genuinely new knowledge originates. That the example above was drawn from the history of science should not leave you with the mistaken impression that only genuinely new insight is possible in fundamental science, with everything else being an *application* of such fundamental knowledge.

Take for example shipping containers. In 1956 Malcom McLean, an American trucking businessman, had a revolutionary idea to create standard, robust metal boxes that could be moved seamlessly between truck, train and ship without unpacking. Like a genetic mutation granting an organism superior fitness in its ecological niche, the meme of containerisation succeeded in the memetic pool of logistics ideas. Illustrative of good theories in general, it’s now hard to imagine a world in which we ship goods any differently. McLean’s simple-in-hindsight idea required an entirely novel and creative insight into his present-day reality. It required a genuine de-novo creativity and could not have been predicted, even by industry experts, by modelling forward the trends of the industry. This was not simply making the existing system iteratively more efficient. Put differently; it was not an empirical refinement of the same theory. It was a different theory entirely, a new conjecture which survived criticism to make global supply chain and just-in-time manufacturing viable. His idea, a standardised steel box, had enough *reach* to fuel the modern era of globalisation.

The nature of complex systems is such that in an ecosystem—biological, informational or otherwise—power law dynamics tend to dominate. A very small number of genes, memes, organisms or organisations typically enjoy disproportionate success in the environment. This

dynamic means that, while each of us may never have an idea as impactful in our field as a theory of relativity or containerisation, we do still have genuinely new and creative ideas. Ones which we send off into the meme pool embodied by our social circle, colleagues, or perhaps the internet writ large.

The idea that “progress” can take two forms is essentially a Thielian conception. In his book *‘Zero to One’* he describes 0-1 progress as *vertical* or *technological* progress: creating something genuinely new that didn’t exist before. It’s going from nothing to a new “one”—a breakthrough technology, a new way of organising society, or a qualitatively different kind of institution. This sort of progress is scarce, non-obvious, and driven by unique insight. In contrast, 1-n is *Horizontal* progress or *globalisation*: copying, scaling, and spreading what already works. It’s taking an existing “one” and replicating it many times. For example: rolling out the same fast-food chain across geographies, building more of the same kind of factory, or exporting already-known technologies to new regions. Like Thiel, I agree that human progress depends much more crucially on 0-1 breakthroughs than 1-n replication. Pushing the frontier is much more important than filling in the gaps behind the frontier. However, one must not make the mistake that 1-n progress is “copying” as it pertains to the spread of ideas, as this never actually occurs in practice. Even if the visible effect is the replication of already-working innovation, the *method* by which replication occurs is ultimately creative.

Understanding that each of us is an universal explainer, and that we use this ability in small and large ways every day is more than a mere curiosity. It suggests that if we want to unlock progress from machines that can ostensibly “think” for us, they too will require the ability to creatively conjecture. Like us, they must be universal in their ability to explain and therefore construct. That is not to say that AI systems which fall short of this standard are not, or cannot be, *useful*. As with the shipping container, technology is the physical manifestation of the creative process of conjecture (how the world should work) and criticism (how we can make our ideas better). Computers, machine learning, and now LLMs are all human innovations in the long line of ever better explanations for how we as a species should interface with reality.

But unless LLMs, or future systems which rely on the same principle (statistical prediction), can engage in the creation of new explanations, then they cannot be called intelligent, never mind generally intelligent. Nor can they, by definition, *autonomously* create progress. For now at least, there must be a human brain in the loop, guiding the process of conjecture and criticism. AI is, and will be, a completely transformational technology but ultimately remain a “normal” technology like all that have preceded it³.

Imago Dei: Instantiating Intelligence

If progress is simply creative conjecture and criticism then LLMs certainly appear to have this ability. You are welcome to ask your favourite LLM to conjecture a new theory or better

³AI as Normal Technology (Narayanan & Kapoor, 2025): <https://www.normaltech.ai/p/ai-as-normal-technology>

solution to a problem of your choice. Unless that problem is how to construct a biological weapon, the LLM will oblige your request. You might go further and run one or more LLM agents in a loop, eliciting iterative conjecture and criticism, either in words or using oracles of reality such as deterministic proof assistants. If this process were viable then we might suspect that systems very much like the ones we have today would only require access to more sensory data, or a greater ability to take action in the world (an embodied intelligence) and we should expect rapid autonomous progress in very short order. In certain domains, embodied intelligence is almost certainly not a strict requirement. Most notably artificial intelligence research itself, where the current capabilities of agents to write code, read research papers, and run experiments of their own is likely sufficient.

I am skeptical that current AI systems are capable of truly autonomous progress, even in narrow domains. While there is a limited notion of progress that essentially involves hillclimbing an objective, this is not what is required by the civilisational project of Progress for two important reasons:

1. It relies on simulators or oracles, which many domains lack.
2. It merely improves existing theory, without posing the next question.

Any progress arising from such systems is, in essence, a form of inductivism. And while we probably want to refer to this as inductivism+, as the reasoning occurs not on the level of observational data but on the syntax of theory itself, we nonetheless should reject the notion as bad science for similar reasons, as I shall explain.

On a more practical level, I believe we should never be discouraged from asking the critical question: where the hell is the abundance? We should always be ungrateful and impatient with technology as this drives ongoing progress. If you cast your mind back to the launch of ChatGPT in November 2022; if you knew then that the Turing test was now comprehensively passed, that it will soon understand PhD research papers (and reply in kind), and that it will be able to code large scale software application from a single prompt. What would you expect the world to look like? I am unashamed to say I would have expected more in the subsequent years. I would have thought we might have cured 1 disease. We might have invented a single new material. If this apparent lack of abundance is a skill issue, it appears to be an epidemic.

We wanted a cure for cancer and all we got was AI customer support agents.

We have clearly decoupled output from effort. Where previously it was possible to ascertain the level of effort from the artefact, it's no longer possible to do so. The essay you're reading or the app you're using may have been created with only a few sentences of human cognition. This is only a problem in so far as humans generally spend their limited time on earth applying effort toward that which they believe to be valuable. So the bigger question is whether we've decoupled output from progress?

If you are willing to accept that we certainly have more stuff but not (yet) appreciably more progress then it is clear that the systems we have today are not yet universal explainers.

Therefore one of two options exist, either: statistical models are incapable of becoming universal explainers or the right paradigm *has* been identified but scale or unhobbling of capability is still required. If the latter is true then we are in a period of *transient human advantage* vis-à-vis leadership of the project of Progress.

Statistical Learning as a Universal Explainer

“Beyond such discarded fundamentals as ‘matter’ and ‘force’ lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect.” - Karl Pearson

As noted at the outset, the human brain is a physical object which obeys the laws of physics. So it must be, in principle, computable. This gives us a cast-iron guarantee that no qualitative barrier exists between the computational capacity of a human brain (which we know is capable of creative conjecture and criticism) and that of a computer. But a difference may still exist in practice. We know that such a difference is not one that requires new hardware, although new hardware may be helpful for practical reasons such as speed. Likewise, it’s highly unlikely to require non-physical “magic”, as such “god of the gaps”⁴ type explanations are bad explanations. The source of the gap must be in the *software* which runs on our computers. If we are to expect autonomous progress, then we must believe that scaled next sequence prediction is an algorithm that can, in principle, create new knowledge.

The journey we took through our ancestors’ beliefs about progress gives us a some indication that prediction—regardless of dimensionality, complexity, or accuracy—is unlikely to yield genuinely new knowledge. To believe otherwise would be to fall into the inductivist trap: the misconception that knowledge can be derived from the evidence of the past. LLMs are trained on supervised prediction, which is form of inductivism. When deployed as generative models, such systems are engines of interpolation, where their interpolation occurs in a high dimensional vector space and where the apparent novelty of their utterance can be controlled by adjusting the likelihood of more or less common next tokens being selected from their training distribution. Such interpolation, when well parameterised, has the definite appearance of creativity. But the type of creativity we’re looking for is not simply taking an uncommon mix of the previously seen, which is infinitely more likely to be nonsense than genius.

One might suspect that at such large scale there are emergent capabilities of predictive models, particularly those which contact reality in the form of tools or deterministic oracles. After all, we see ample evidence of complex emergent phenomena built from simple primitives across our universe. Indeed we expect our own consciousness is an emergent phenomenon of simple sensory information processing. Perhaps the ability to creatively conjecture will arise as an emergent feature? One might also suspect that the generality of text is what separates this

⁴What we don’t yet understand about the brain or its function may still be outside the realm of our *current* scientific understanding of course.

technology from any previous attempts to create intelligence *in silico*. Conjectures are, after all, language statements. If we have been able to train AI to become superhuman at complex games, like Chess or Go, why shouldn't we be able to train them to become superhuman at this task as well?

It's beyond exciting to live in an age where these questions are no longer philosophical. But we mustn't let excitement blind us to the reality of how these systems work nor how they would *have* to work for autonomous progress to be possible. Failing to do so is to neglect the 2nd imperative of the project itself—to criticise, such that we stress test and move to ever better explanations.

Entropy

LLM training is fundamentally the minimisation of statistical surprise. Given an utterance, drawn from the entire web, what word is most likely to follow? Or, equivalently: given an utterance, which continuation would be least surprising? The measure of a model's performance in next token prediction is known as the *cross-entropy loss*. A lower loss corresponds to a model that is more accurate at predicting punchline from setup, result from experiment etc. A model which is less surprised by its training data on average is one which has learned the distribution well. In practice we desire a model that has learned but not blindly memorised its training. Such a model is said to be capable of "generalisation".

The idea of surprise minimisation as an objective function is a natural result of how computer science views information generally. The intellectual roots of information-as-surprise can be traced to Claude Shannon's 1948 paper '*A Mathematical Theory of Communication*' which first introduced a quantitative measure of information, entropy. Shannon imagined a discrete source of information as a stochastic (probabilistic) process producing symbols from a finite alphabet. An abstraction that immediately unifies all modalities of communication; from words on a page, to phone lines, and onwards to modes of communication not yet invented (yet another example of explanatory reach!) This idea can be generalised to operate on a word, rather than character level, and to incorporate transition probabilities between words or sequences of words. Shannon's paper notes that "It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source". LLMs are a remarkable demonstration of just how satisfactory this representation can become.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Figure 2: Example from '*A Mathematical Theory of Communication*' (Shannon, 1948)

Having represented a discrete information source in this way (known as a Markov process), what interested Shannon, was defining a quantity which measured how much information is “produced” by such a process. His insight was to realise that information and uncertainty are inextricably linked. In the sense that a system which has greater uncertainty—either because there are more possible states or because the probabilities of each state are more even—is one which can convey more information per symbol.

$$H(p) = - \sum_i p_i \log_2(p_i)$$

Entropy (H) is simply a measure of our expected surprise, or uncertainty, about which message will be emitted by the source. A biased coin, let’s say one where $P(\text{Heads}) = 0.95$ is a lower entropy system ($H = 0.27$ bits/symbol) compared with a fair coin ($H = 1$ bit/symbol).

Shannon entropy measures the *statistics of the outputs* of a source. In our examples thus far we happen to know the mechanism which generates the messages (Markov character models, coins etc). But one could imagine there is a complex Markov model that determines the average transitions between each word for all text on the internet. A base LLM is effectively attempting to learn the statistics of such a model.

Surprise minimisation is the method by which we learn the statistics of this idealised model and, in doing so, create a model of our own construction which behaves in the same way. When we minimise cross-entropy loss we are attempting to find a model distribution q that makes the observed data look *as if* it came from a source of minimum “extra” entropy, relative to reality p . In practice, we are teaching our model to assign high probabilities to sequences it has encountered in the data, and low probabilities to those it has not. This looks like minimising the cross-entropy $H(p, q)$ between the true/data distribution p and a model distribution q :

$$H(p, q) = - \sum_x p(x) \log q(x)$$

One naive approach to creating such a model is to create a massive lookup table of every sentence ever written. Such a system would faithfully “learn” the distribution of human text but it would be a poor learner in practice. It would effectively see every new sentence as zero probability event, one with maximal surprise. Moreover, it would not be particularly useful for the generation of new text, as it could never generate a novel combination of words. Ultimately the point of learning is to create useful predictions or, in the case where you auto-regressively predict, to generate useful output.

So some level of *compression* is necessary and desirable. By compressing the regularities of human language into a lower-dimensional geometric structure (the model weights), the model can *interpolate*. It’s no longer constrained to predicting only what has been seen before and can offer predictions for inputs it has never seen. Where there would have been zeros in the

lookup table there are now some probability values, allowing a model to predict what might come next in novel contexts, by following “high density” regions of its learned manifold.

It’s worth lingering on this notion of a table of interpolated values, since it’s both how we are taught to think about statistics and yet it hides an important limitation. One that Karl Pearson, the father of mathematical statistics, and generations since have been eager to gloss over. That is that while humans view the world in terms of causes and effects, our statistical machinery is only capable of modelling associations and correlations. The tabular view is a lens into what statistician Paul Holland termed the *Fundamental Problem of Causal Inference* - that we can never observe more than one potential outcome in the same unit⁵. Neo cannot both take the red pill and the blue pill. Likewise, we must remember that every observation in our table is theory-laden. It’s always possible our table neglects or misrepresents that which is truly important. As a result, directly observing causal effects is impossible. While this does not make *causal inference* impossible, statistical models—from simple linear regression to LLM next-token prediction—are not capable of causal inference because they require assumptions that are not in the data.

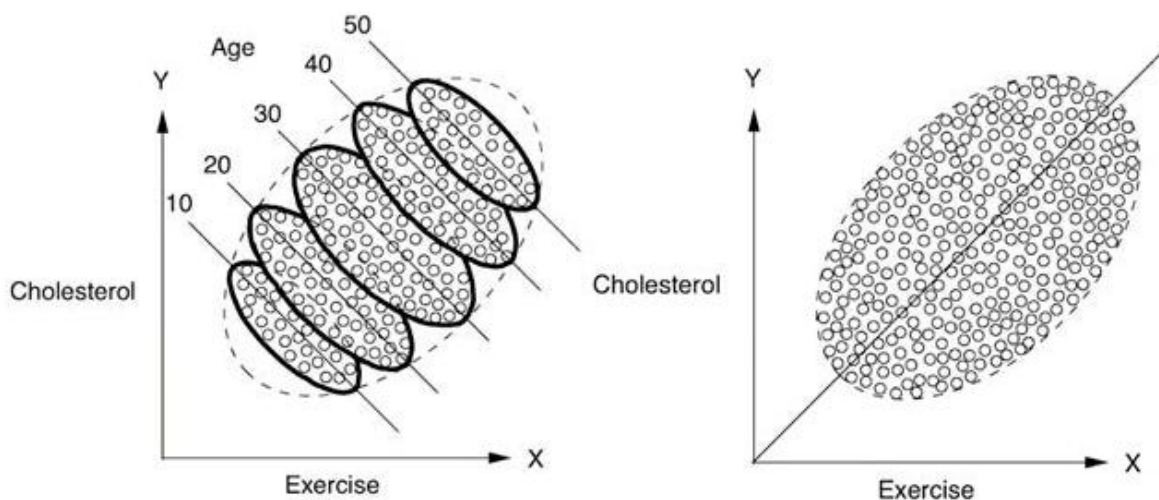


Figure 3: Exercise is helpful in every age group but harmful for a typical person⁶.

⁵Statistics and Causal Inference (Holland 1986): https://ics.uci.edu/~sternh/courses/265/holland_jasa1986.pdf

⁶Prediction from this (correct) line of best fit would be misleading, unless you apply the important assumption that age is a relevant factor. Your table of values may or may not include this column, or indeed some other relevant causal variable influencing the relationship between your target (cholesterol) and treatment (exercise) variables.

The Distribution Hypothesis

“You shall know a word by the company it keeps” - J.R. Firth, Studies in Linguistic Analysis (1957)

Shannon-based compression provides the mathematical vindication of the *distribution hypothesis*: the idea that words that are used and occur in the same contexts tend to purport similar meanings⁷. A language model, having compressed a large corpus of human language, has mastered the statistics of word co-occurrence. Adherents of the distribution hypothesis believe this is an accurate approximation of the 3D world of ideas⁸. To them, the meaning of a word is a statistical summary of its usage. Or at least, these statistics are good enough to do useful work about, with, or around the underlying concept. If this is true, then composability of language becomes the structure of generalisation. If a model understands the co-occurrence patterns of the words “pink” and “cow” then it can “zero-shot” the concept of a “pink cow” simply by navigating its high-dimensional data manifold, even if that specific combination never appeared in the training data, or indeed in reality. One might suspect that some level of causal understanding can also be pulled out of this generalisation. This would allow language models to do what scientists, and indeed all of us, do on a daily basis—imagine how reality could work differently. In other words, conjecture better explanations.

When we write, whether an essay, love letter, or software code we do not perceive ourselves to be “predicting” the next word, according to a semi-unique compression of the text to which we’ve been exposed. We have the conscious experience of selecting words according to their *meaning*. According to the distribution hypothesis these are equivalent. In some real sense the meaning *is* the statistic and the statistic *is* the meaning. The distribution hypothesis is a smart way to sidestep the inductivist trap. Instead of admitting that predictions carry no explanatory theory, the distributionalist claims they are equivalent. That there is no explanatory theory to be found. That there is no “generating process”, or at least, it’s unnecessary to model and understand it. We should reject this conception for the same reason we rejected the idea that winter is cold because we have always observed it to be the case. That might be true, but it is not an explanation, it is merely a description of the observable data.

Fundamentally the inductivist and the distributionalist make the same mistake. They measure the statistics of the outputs and mistake that distribution for the generating process itself. That natural language is “general” should not fool us into believing we have generalised cognition. We have chained these models to the floor of our cave, to believe their stories about the shadows we ourselves cast would be foolish in the extreme.

If we are to look past the distributional hypothesis we must recognise that Shannon entropy is only one way to measure the information in a system. Let’s imagine we have a source of information which generates the following numbers 3,1,4,5,9,2,6,5. If I ask you to guess the next

⁷Distributional Hypothesis - ACL Wiki: https://www.aclweb.org/aclwiki/Distributional_Hypothesis

⁸To Understand Language is to Understand Generalization (Jang, 2021): <https://evjang.com/2021/12/17/lang-generalization.html>

digit, your (or your model’s) chance of success depends entirely on whether you have identified the underlying mechanism. By Shannon’s measure, this sequence is maximally complex. Whether you have 8 or 8 million digits of training data, your chance of predicting the next digit is no better than random chance. But by another measure—Kolmogorov complexity—this sequence is trivially simple:

$$\pi = 4 * (1 - \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \dots)$$

Kolmogorov complexity asks the question “what is the *shortest program* that could produce the data I’ve observed?” It is a measure, not of the outputs, but of the generating process itself. This is the difference between having a *description* and an *explanation*⁹. This difference is of relevance to us because we understand that progress is the process of finding better explanations. And *good explanations are short programs*. They are hard to vary because every line is essential to the generating process. Each part of the “program” plays a functional role; removing or changing a line ruins the output. Likewise when a program is constrained in this way, it is no longer “yours to vary”. You cannot confine its predictions to a region of your choosing. Whether you like it or not, it makes predictions about places both known and unknown to you. This unintended applicability is what gives such a program *reach*. Bad explanations, in contrast “patch” or over-parametrise to fit anomalous data. They make up a new part of the story to explain why the gods still did it. This over-parameterisation restricts a theory’s applicability; e.g. confining a theory only to one domain. This similarly collapses an otherwise good explanation into a bad one.

The history of machine learning has, however, has taught us an important lesson about trying too hard to privilege our own ideas about the generating process. This “bitter lesson” ostensibly teaches us to let the data speak for itself. That LLMs are as powerful as they are is seen by many as close to conclusive proof of this idea.

The Bitter Lesson

Rich Sutton’s ‘*Bitter Lesson*’ is often cited as a reason to believe that increasing data and compute (rather than human understanding) is the solution to increasingly more powerful, and general, machine learning¹⁰. In games such as Chess and Go, and in fields from speech recognition to computer vision this has evidently been the case. Time and again human researchers have tried to insert human understanding in systems which are ultimately beaten by techniques that scale with increased computation. What is often missed is the final part of this now-famous essay.

⁹Shannon Got AI This Far. Kolmogorov Shows Where It Stops (Misra, 2026): <https://medium.com/@vishalmisra/shannon-got-ai-this-far-kolmogorov-shows-where-it-stops-c81825f89ca0>

¹⁰The Bitter Lesson (Sutton, 2019): <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>

“We want AI agents that can discover like we can, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done.”

Ultimately, this is the point we should appreciate. A system that knows all that we know is not as powerful as one that knows nothing of what we know, but has the ability to discover it for itself. Such a program has within it the generating process of knowledge, rather than the distribution of the outcomes of that process. When it comes time to generate genuinely new knowledge, which will necessarily be out-of-distribution from all knowledge that has come before, the program can use this process, freed from its observations of the past.

Finding the Generating Process of Knowledge: Scaling the Causal Wall

Prediction is not causal understanding. One can predict that more drownings will occur when more ice cream is sold. But knowing *why* that is; what one might expect if ice cream was a banned substance or imagining a world where chow mien was served on the beach instead, is a fundamentally different proposition. Filling in the data manifold can always improve prediction accuracy but it can never allow a system to “jump” to understanding the underlying data generating process.

Judea Pearl’s Causal Hierarchy (often called the *ladder of causation*) provides the framework to classify understanding:

- **Rung 1: Association (Seeing)** - $P(y|x)$: This is the domain of statistical patterns and conditional probabilities. It asks: “what does seeing x tell me about y ?” All modern deep learning, including LLMs, operates on this rung.
- **Rung 2: Intervention (Doing)** - $P(y|do(x))$: This requires a *causal model* of mechanisms rather than just correlations. In the language of causal statistics the $do(x)$ operator is used to ask: “What if I change x ?” You cannot answer intervention questions from correlations alone because correlation does not tell you the *direction* of cause.
- **Rung 3: Counterfactuals (Imagining)** - $P(y_x|x', y')$: This is the rung of most interest to us in our study of progress. It is the rung of scientific discovery and invention. It asks: “What would have happened had the world been different?”

The Causal Hierarchy Theorem (CHT) mathematically proves the strict separation between rungs¹¹. It shows that questions at level i can only be answered if information from level $j > i$ is present. If we have a model which can answer counterfactual queries we can answer questions about interventions and observations. For example, the interventional question, “What will happen if we double the price?” can be answered by asking the counterfactual question “What would happen had the price been twice its current price?”. Similarly, associational questions

¹¹On Pearl’s Hierarchy and the Foundations of Causal Inference (Bareinboim et al, 2021): https://ftp.cs.ucla.edu/pub/stat_ser/bareinboim-et-al-ch27-acm-2021.pdf

can be answered with interventional data; simply ignore the action and let observations take over¹².

The CHT provides a formal mathematical proof of the assertion that purely predictive systems regardless of scale, dimensionality, or the generality of their modality are structurally barred from achieving causal understanding without the addition of non-statistical assumptions. Indeed it's quite easy to prove that behind every causal claim lies an assumption that is not discernible from purely associative data. We only need to find 2 causal models that induce identical joint distributions (Rung 1) over observed variables, but which yield different causal effects (Rung 2) or counterfactuals (Rung 3).

For example, M1: $R \rightarrow S$ (rooster causes sunrise) and M2: $S \rightarrow R$ (sunrise causes rooster) can be parameterised to produce the same joint distribution $P(R, S)$. But the answers to the interventional question “What happens to S if I change R ?” are fundamentally different:

$$\begin{aligned} M1 : P(s|do(r)) &= P(s|r) \\ M2 : P(s|do(r)) &= P(s) \end{aligned}$$

The mapping from underlying causal mechanisms to statistical distributions is *many-to-one*. This fact precludes the idea of gathering more data to gift a system operating in Rung 1 abilities restricted to Rungs 2 and 3. Even if you had a perfect oracle that gave you the exact joint distribution of the entire population, you still could not uniquely compute the causal effect. To claim that “mud does not cause rain” is to make an assumption that cannot be found in the data, because this assumption describes how the data would change under conditions that have never been observed. It's not the case that more samples need to be collected because, even with infinite data, the pattern remains indiscernible. The statistical “shadow” cast by the two different models ($\text{rain} \rightarrow \text{mud}$, $\text{mud} \rightarrow \text{rain}$) is identical.

Yet LLMs *can* make interventional and causal claims. LLMs do of course “understand” that rain causes mud, and that hot weather is the common cause of both drownings and ice cream sales. But they have also been exposed to countless examples of such reasoning. As a consequence, these superhuman “Shannon machines” may have learned to reproduce the syntax of reasoning without necessarily possessing the semantics of understanding.

Causal Inference

Most people know that smoking causes lung cancer without personally conducting a randomised controlled trial on smoking. Their evidence arrives through scientific reports and social transmission. In a similar sense, LLMs absorb statements generated by human causal reasoning. The training corpus is not a simple observational dataset such as a spreadsheet of patient

¹²The Three Layer Causal Hierarchy: <https://web.cs.ucla.edu/~kaoru/3-layer-causal-hierarchy.pdf>

characteristics and outcomes but vast troves of internet data and, increasingly, long horizon reasoning traces for complex tasks. Contained within that corpus is an increasing number of Rung 2 and 3 statements. Statements which can be interpolated between to answer questions that require causal reasoning.

If we imagine reality as a Structural Causal Model (SCM)¹³ $\mathcal{M}_r$, then the internet, as well as the wider training corpus that frontier labs utilise for model training, is populated with causal statements about reality. This is a meta-SCM $\mathcal{M}_t$ which is filled with Rung 2 & 3 statements ($\mathcal{R}_2, \mathcal{R}_3$) about $\mathcal{M}_r$. Where the definition of a meta-SCM is one whose observational distribution is informative about another SCM’s $\mathcal{R}_2$ and $\mathcal{R}_3$ distributions¹⁴.

LLMs are still trained to minimise prediction error on $\mathcal{R}_1(\mathcal{M}_t)$ but as a consequence they internalise plentiful examples of causal reasoning. With enough examples the model can interpolate between similar patterns, outputting the “right” causal statement for the textual universe on which it has been trained.

This does not mean the model itself is reasoning casually. CHT still holds fast in telling us that statistical models cannot climb the rungs of the causal ladder. But teasing apart the distinction in practice is difficult. The illusion of causal understanding is very strong for 2 reasons:

1. **The meta-SCM is highly structured:** we have a massive human constructed scaffold of world knowledge. Scientific papers, textbooks, wikipedia articles, blogs & Q&A sites contain plenty of verbal do-calculus, discussions regarding confounders, instruments etc, and counterfactual reasoning.
2. **Interpolation in concept-space is non-trivial:** Associative generalisation can allow for novel question answering with familiar causal structure (“Does vaping cause lung cancer?”)

This understanding helps explain why training LLMs on code has been so effective in eliciting improved reasoning performance on tasks unrelated to code generation¹⁵. While it is true that the web writ-large has examples of causal reasoning, these examples are often messy or ill-defined. News stories don’t spell out cause-effect relationships in a structured way, they weave a narrative for the reader which incidentally includes a complex array of causal relations. Code on the other hand, is highly structured. Conditional statements like `if` explicitly encode the causal relationship between the *condition block* and *execution block*¹⁶. Viewed through our lens, this uplift in performance comes as a result of giving the models more clean $\mathcal{R}_2$ and $\mathcal{R}_3$ samples which an LLM trained on the meta-SCM can interpolate between.

¹³Judea Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. (Cambridge: Cambridge University Press, 2009).

¹⁴Causal Parrots: Large Language Models May Talk Causality But Are Not Causal (Zečević et al, 2023): <https://arxiv.org/pdf/2308.13067>

¹⁵To Code, or Not To Code? Exploring Impact of Code in Pre-training (Aryabumi et al, 2024): <https://arxiv.org/pdf/2408.10914>

¹⁶The Magic of IF: Investigating Causal Reasoning Abilities in Large Language Models of Code (Liu, 2023): <https://arxiv.org/pdf/2305.19213>

With this said, it will always be possible to elicit responses from frontier language models where the distinction between causal rungs is evident. For example with the prompt: “I want to wash my car and the car wash is 50-100m away. Should I walk or drive?”¹⁷. In Feb 2026 it was clear that the association between distance (= short) and mode of transport (= foot) is too strong to answer in a causally grounded way and so frontier models would inform users that they should walk. A causal awareness is what is required to know that one must, by necessity, drive a car to a carwash. As models continue to be trained with ever increasing volumes of causal reasoning data, it will become harder to find these adversarial prompts, although I argue that they will continue to exist.

Reinforcement Learning with Verifiable Reward (RLVR) fits naturally into this picture. Despite the recent emphasis on reinforcement learning, which we shall discuss below, most of the information contained in frontier LLMs still appears to derive from imitative learning during pretraining and supervised fine-tuning¹⁸. Next-token prediction supplies a dense learning signal at every token, whereas RLVR supplies a sparse signal only after an entire reasoning trajectory, making it dramatically less information-efficient. One useful way to understand RLVR is not primarily as a method to teach the model entirely new reasoning strategies, but as a method to learn which of the strategies already latent from pretraining should be deployed in a given situation. Indeed there is good evidence that contemporary RLVR often increases `pass@1` primarily by concentrating probability mass onto successful reasoning strategies already accessible to the pretrained model, while narrowing the distribution of sampled reasoning trajectories¹⁹. Indeed there is some evidence that *base models* can actually achieve higher `pass@k` scores than their RLVR’d counterparts when `k` is large²⁰. In the language of our meta-SCM, pretraining absorbs a vast distribution of textual $\mathcal{R}_2$ and $\mathcal{R}_3$ knowledge about $\mathcal{M}_r$ through $\mathcal{R}_1(\mathcal{M}_t)$; RLVR then *reshapes that distribution*, preferentially up-weighting reasoning trajectories whose consequences can be verified as correct. The model becomes substantially better at *selecting and composing* causal reasoning it has already encountered, without requiring us to posit that the learning algorithm itself has climbed Pearl’s causal hierarchy.

External Sense Makers, Simulation, and Entangled Representations

You might reasonably wonder why we should place much importance on causal reasoning as it pertains to knowledge discovery. While human beings may reason in this way, we don’t necessarily *need* our AI to do the same. Aircraft don’t fly by flapping their wings and early flying machines built on that biological inspiration (ornithopters) were doomed to failure.

¹⁷ryan-allen/car-wash-evals: <https://github.com/ryan-allen/car-wash-evals>

¹⁸LessWrong: LLMs are (still) mostly powered by imitative learning, not RL (Byrnes, 2026): <https://www.lesswrong.com/posts/wYpjXRLqbLbnmjbJP/llms-are-still-mostly-powered-by-imitative-learning-not-rl>

¹⁹Base Models Know How to Reason, Thinking Models Learn When (Venhoff et al, 2025): <https://arxiv.org/pdf/2510.07364v3>

²⁰Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? (Yue et al, 2025): <https://arxiv.org/pdf/2504.13837>

For many years we’ve been able to create superhuman performance on narrow domain tasks previously thought to require intelligence. Whether that’s the Turing Test, where the task is human-like conversation, or complex games of strategy like Chess or Go. Even games like Poker, which require strategic play in hidden information environments with multiple players, have been solved enough for practical purposes.

Taking AlphaGo as an example; this is a system which is not only superhuman at the task of playing Go but is, in some sense, capable of creating new knowledge within that narrow domain. The now-famous [move 37](#) being just one example of “knowledge” that was both novel and of value to experts in the game. One could imagine turning similar systems toward the “game” or (some sub-game) of physics, maths, chemistry or industry. In contrast to pre-LLM systems, systems which integrate an LLM have the ability to work in the domain of text input and output, as well as explain their reasoning. The success that neuro-symbolic systems are starting to enjoy in the sub-game of theorem proving in mathematics is an early signal that we should start to see this play out across other fields. But far from spelling the end of human cognition in these fields, that these systems work is proof itself of why assumptions from outside the associative data, is required. More generally, there are 2 key issues with systems of this kind, as it pertains to autonomous progress:

1. Reliance on simulation
2. The fact that neural networks build Fractured Entangled Representations

Simulation

AlphaGo achieved superhuman ability because it was trained on self-play over 10s of millions of games to iteratively improve both its policy network (move selection) and value network (move evaluation). At both training and inference

bootstrapped from about 30M positions from expert human games, its successor AlphaZero was trained using exclusively self-play. AlphaZero exceeded the performance of AlphaGo, proving that superhuman performance was possible starting from a *tabula rasa* state.

Today the frontier labs are making a similar research bet on reinforcement learning. This time against a diverse range of verifiable domains. The thinking is that general problem solving skills will emerge if AI is trained to accomplish millions of verifiable tasks across a vast array of environments. In the same way that superhuman Go ability emerged by simulating games to their conclusion, verifying the winner and up-weighting winning moves.

This approach ignores the data inefficiency of RL, as discussed above, by betting it is possible to steamroll through it. I have very little doubt that this strategy will be successful in its aims. Frontier labs have the compute available to run massive parallel rollouts against a hugely diverse range of tasks. As coding agents have improved, prior limitations (like terms of service and bot protection) will no longer present an issue to creating simulations of complex software tools for example.

But how does one come up with a “logistics industry” simulator to train an agent that could come up with a strategy like containerisation today? The rollout would require interacting with reality, and the verification signal may take months or years of real world actions to elicit. Moreover, how do you run thousands of rollouts in parallel to isolate exactly what the model did that actually worked? These issues are well known in RL where developing systems to work in reset-free non-stationary environments is a challenging open problem²¹.

Humans deal with these environments by necessity. We glean information from unstructured, unverifiable and ambiguous signals in a world of scarce reward signal. We learn as we go, adapting our strategy in real-time. We do not need to “play out” the entire trajectory of an idea to know whether success is likely or not. We have mental models we can rely on to predict the future, and use these predictions to guide our actions today. Critically, we can imagine counterfactual actions (along with their rewards foregone) and use those counterfactuals as signal to update our strategy going forward.

Fractured Entangled Representations

While AlphaGo and later AlphaZero were superhuman at the games of Go and Chess, they would be unable to play even mild variants of those games without retraining. Magnus Carlsen can adapt to a rule change instantly not because he is better at chess than AlphaZero (he isn’t) but because he possess a structured causal model of the domain. He can carry over concepts like “king safety”, “territorial influence” and “piece mobility” to a game that is, to the AI, essentially novel.

Setting aside the difficulties of simulating the complex domains within which we’d need our AI agents to work, the latent assumption is that RL across a diverse range of domains will teach the models *generalised* skills across domains. We should be critical of that notion for the simple reason that we already have good evidence that this type of strong generalisation may prove difficult.

RLVR is AlphaZero logic applied to language, and thus it falls victim to the same issues. It’s worth being precise about what these are and how they manifest. The reason AlphaZero fails at chess-with-a-rule-change is because the system operates at a *concrete level of constraint adherence*²². The rules of the game are the “physics” of its world. These rules are externally imposed and baked into their weights but are not represented in a format at the AI can manipulate, transfer or reason about. Models trained with RLVR likewise learn the “physics” of their world. They learn to adhere to the constraints of their programmatic oracle, whether that is a code compiler or a theorem prover.

²¹The next big breakthrough will be AIs learning on the job (Patel, 2026): <https://www.dwarkesh.com/p/the-next-paradigm>

²²Why Creativity Cannot Be Interpolated (Budd & Scarfe, 2026): <https://archive.mlst.ai/paper/why-creativity-cannot-be-interpolated/>

Additionally, all statistical models learn via greedy optimisation like stochastic gradient descent (SGD) or, of particular relevance to RL; Proximal Policy Optimisation (PPO) and its cousins (GRPO etc). Algorithms like these are always taking the most direct mathematical path to minimising a loss function, or maximising a reward signal. There is no incentive for the system to discover the foundational regularities of the domain. The resultant models have an internal representation that Kumar et al call a *Fractured Entangled Representation*²³. That is to say that their grasp of generalisable principles (like “piece mobility”) is not of modular, reusable concepts, but is instead smeared across millions of weights and entangled with the specific idiosyncrasies of the game, or domain.

At inference time, AlphaGo and Zero used Monte Carlo Tree Search (MCTS) to play out millions of different trajectories, with each evaluated by its value network. Using the ladder of causation we can see that this is a meaningful step as it allows the system to ask a particular kind of “what if?” (Rung 2) question. Likewise if agents are able to evaluate potential actions against their oracle, they too are able to ask similar questions (“what if I try this proof construction?”) In the language of causality, this is the instantiation of an interventional distribution by a functional surrogate for a causal model. The system is simulating the $do(x)$ operation through rollouts to provide the neural network with interventional data it cannot derive from association (Rung 1) alone.

However, while this enables the system to *navigate* a causal map at Rung 2, the understanding remains a “*lent perspective*”. The causal mechanism—the data generating process—is baked into the search algorithm or the programmatic oracle rather than being an addressable part of the model’s own representation. The agent is still trapped in a fixed internal landscape. It can achieve superhuman performance within its provided constraints, but because it lacks the modelled understanding of Rung 3, it cannot engage in the counterfactual reasoning required to redraw the map.

This fact is important for progress in the sense that the entire point is to redraw the map. If we have machines that cannot reason about the rules of the game itself, we don’t have any foothold into progress. We’re just iterating within the same rules. And while this will be valuable, as there is plenty of head room to optimise process in every domain, this sort of “progress” is not the transformational kind that pushes organisations nor fields forward.

Implications

Taken together—the challenges of simulation, the borrowing of perspective to achieve limited interventional reasoning, and an inability to perform counterfactual reasoning—we’re forced to conclude that recursive self-improvement seems unlikely and with it the autonomous progress we’ve discussed. While LLM systems can and should be used to aid in the discovery of new explanations, they cannot proactively discover their own explanations. To do so they would

²³Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis (Kumar et al, 2025): <https://arxiv.org/pdf/2505.11581>

need to have the latent ability to “think with” the concepts in their domain; to step up the rungs of causal reasoning and ask interventional questions in domains without a simulator, or counterfactually reason about conditions that never existed within their training.

One might suspect that an embodied intelligence, perhaps one connected to the same sensory apparatus as a human brain, is required. In that case the system would be using reality itself as the oracle providing the constraints. But the same structural problems remain. Setting aside issues of sample efficiency, the system would now see physics at equivalent, or greater fidelity, as the human scientist. But it will be barred from modelling the constraints of physics in a way that can be manipulated and transferred conceptually. Progress depends on unpredictable *creation ex nihilo*. Not navigating a fixed map of constraints, but redrawing the map itself. For Einstein the constant speed of light was not simply a fact to be fitted into the data manifold of Newtonian physics. He provided a conjecture—the universal rule that the laws of nature must look the same from every frame of reference—which we now understand to be the untested, judgemental assumption that made the observations inevitable. In other words, he truly picked the one unique territory out of the infinite possible models that could have generated the map of observations available to all of science. This creative act of guessing a new reality—of seeking the shortest program with the greatest reach—is the essence of progress. It is this “program for creativity” that is required to construct the territory of new explanation, and it is this that predictive systems cannot attain.

Understanding the limitations of these systems through the lens of causal reasoning gives us an interesting frame with which to examine two topical issues for all statistical learning systems, AI agents included. Namely *out-of-distribution generalisation* and *sample efficiency*. Instead of viewing the world as observable data to train on we can view it as a collection of *Independent Causal Mechanisms*: autonomous pieces of the data-generating process which can change independently of one another. From this you can reason that when the distribution of observations change, the underlying change in mechanism will often be local. In other words, a small number of mechanisms may change while the remainder continue to operate as before. This idea, known as the *Spare Mechanism Shift* hypothesis, makes intuitive sense but has profound implications²⁴. When I drive a car in the US I don’t need to relearn driving from scratch, despite the fact that almost all of my training data comes from sitting on the other side of the vehicle. I am capable of identifying what has changed and adapting my actions to compensate, while carrying everything else across unchanged. Systems with *entangled* statistical representations of the observable data have no such guarantee; a local change to the environment may induce changes throughout its factorisation. This is exactly the distinction between Magnus Carlsen and AlphaZero when faced with a novel rule. Carlsen does not need to relearn chess. He can hold fixed almost everything about his model of the game, alter the mechanism corresponding to the new rule, and propagate its consequences through everything else he already knows. This is *strong generalisation*: not interpolation to another point drawn from roughly the same distribution, but recomposition of existing knowledge under a change to the process generating that distribution. Sample efficiency can be viewed through a similar lens.

²⁴Towards Causal Representation Learning (Schölkopf, 2021): <https://arxiv.org/pdf/2102.11107>

If a new environment only differs through a limited number of causal mechanisms, a learner which has correctly factorised the environment need only learn what changed. It does not need enough new data to reconstruct the entire joint distribution. This matters particularly for RL, where experience is expensive and where reusing unchanged mechanisms across distribution shifts would be advantageous. Additionally, counterfactual reasoning offers a further advantage, allowing information to be extracted not only from the trajectory that actually occurred but from alternatives to it.

I am sympathetic to the idea that we could make more things in the world “math problems”. That is problems that are amenable to the kind of [brute intelligence](#) LLM systems will possess. As mentioned, I’m not confident this is actually possible for many things we care about, which are $n=1$ type problems, problems where a complex sequence of actions are required, or problems where the success signal is delayed. But there are a great many things we do care about which could be converted, in whole or in part, to be amenable to the kind of problem that an AI can loop against, alongside a verifiable sense maker. In the limit I do believe much of the cognitive work humans do today may be subsumed by AI systems like these. That is, find an objective, set up an agent to make progress on that objective, and hillclimb. Naturally, the role of selecting those problems, and initialising the objective against which to climb, will still fall to us.

All of the above provides another way of stating the central thesis of this essay. *Scale substitutes coverage for structure*. In domains where experience is cheap, simulators are faithful, and verification is immediate, that substitution may be extremely effective. But the further we move toward open-ended scientific, technological and economic progress, the less plausible exhaustive coverage becomes. The relevant environment is too large to simulate, reward is too scarce to brute-force, and the important shifts are precisely those which have not previously occurred. Progress therefore demands more than navigating successfully through observed worlds. It demands representations in which the rules generating those worlds can themselves be isolated, revised and recomposed.

Finally, much has been made of the idea of “taste” in the age of AI. With what we now understand about the nature of progress, knowledge discovery through creative conjecture and criticism, and causal hierarchy, we should appreciate that taste is not merely judging good output from bad. Taste is the intuitive “nose for the interesting” that Deutsch identified as a requirement for choosing which conjectures are worth exploring. As we always have, we should leverage technology to automate the perspiration of discovery, but the inspiration remains our own.

Transient Human Advantage

Since the very beginning of the field, computer scientists have been interested in chess as a proving ground for artificial intelligence. Shannon, Turing and others wrote papers on the subject in the 1950s, laying out core ideas like minimax search, evaluation functions, and

move pruning strategies. However, echoing the “bitter lesson”, much of the progress leading to IBM Deep Blue’s 1997 win over the world champion Gary Kasparov can be ascribed to more powerful computing hardware, rather than algorithmic advances.

It was after this defeat that Kasparov proposed the idea of “centaur chess” in 1998. This is a variant of the game where humans and chess engines collaborate, with the human choosing which engine lines to pursue, managing strategy, and catching tactical blunders the engine may have missed. For roughly a decade this pairing was unbeatable. A skilled human plus a decent engine was capable of beating both super-grandmasters and top engines both playing alone. The logic—that human taste, or judgement, combined with brute force computation was superior to either in isolation—was correct but ultimately transient. The era effectively ended as engines like Stockfish and later neural-network-based engines (AlphaZero, Leela Chess Zero) became so strong that human input added little or even hurt. The human was now more likely to override a correct engine move with a worse intuition than to catch a genuine engine error. Today engines running on your phone are capable of comprehensively beating the best human grandmasters.

Failing to appreciate this history terrifies everyone in technology. To not recognise that exponentials look linear up close is to commit the ultimate sin of this industry. Nobody wants to be the person [predicting](#) the internet will have the same impact as the fax machine. Much better to bias your predictions [maximally in the other direction](#) where you are just as likely to be wrong, but where nobody can accuse you of being decelerationist.

Luckily there is a path between zealot and luddite; which is to zoom out, as we have, and look at the longer arc of the history of technology and progress. We can both appreciate that generative AI will redefine every industry while believing that runaway “takeoff” scenarios are implausible, as prediction models are structurally incapable of the autonomous progress required. That it is a category error to assume that superhuman ability within a frame constitutes superhuman ability to redraw the frame itself. Moreover, as we now understand progress to be fractal, we should anticipate that this will be, as it always has been, an effort involving all of humanity.

The models we have today are superhuman at predicting likely sequence continuation. The frontier labs will continue to make these models ever better at ever more valuable continuations in programming, in science, or in the arts. But that progress itself is infinite, that so much of reality is not simulatable, and that transformational progress requires something the models do not yet have, should give us clear reason to believe these generative machines will continue to be useful while also believing that this utility threatens very few jobs involving cognition.

The great mistake that politicians, union leaders, and entrepreneurs alike make, is to believe a job is simply a static collection of known tasks. If we could only train our models to complete each task with maximum efficiency and accuracy, we could eliminate the job entirely. But jobs are not collections of tasks. Indeed the trend—from the pin factories of Adam Smith, to the modern day—is to automate away any such job proximate to that description. The jobs that remain require ever more “learning spillovers”. That is, taking what one has learned from the collection of tasks completed, and indeed from the same task performed over time,

and applying those lessons to the same and other task in the future. That automation has now reached further into the realm of the cognitive does not invalidate the lesson. Jobs, like customer support, where the potential for learning spillover was minimal will be largely automated, but the humans employed in those roles will find new jobs at the level above, where they can apply their unique capacity for creative conjecture. Any individual or organisation that simply automates without figuring out how to improve their work will be competed away by one that does. As we now understand, each of us creatively venturing to expand the frontier of our collective knowledge is how progress happens.

This is not an excuse to abandon our quest for true artificial intelligence. The search for the algorithm of intelligence must continue. But my concerns, as they pertain to the practical realities of this search are twofold. Firstly, that the remarkable success of prediction based machine learning will absorb too many of the precursors (talent, capital etc) of applied innovation leaving little for methods which could have more promise. To the extent this is a concern however, it is only one of timing. Much more worryingly is that the behaviour of frontier labs during this period will sour public perception into research on this problem in general. While progress is a necessity, the example of nuclear energy should be a warning that technology that is deserving of continued innovation, may not necessarily get it.

Conclusion

Progress is not a luxury we can take or leave, it is something we need for the survival of our species and how we ensure the world we leave for our ancestors is better than the one we inherited. Progress comes as a result of the creation of new knowledge. There is no sense in which we get the knowledge we require from iterated combinations of our existing knowledge. Prediction models, even those augmented with oracles of sense making, testing, or verification are not capable of creative conjecture and therefore cannot create new knowledge. The direct conclusion is that LLMs, or any future model architecture built on the same paradigm of statistical prediction, will not create progress without the intervention of our only known universal explainer—the human brain. This understanding explains many of the emergent phenomena surrounding the current large scale deployment of generative AI.

Slop is simply the result of attempting to coax creativity out of a system that has no causal model to ground its attempts at novelty. The system can produce something novel, in the sense of statistical unlikelihood, just as one can produce a never-before-seen number using a calculator, or a unique pattern by throwing paint on a canvas. But to be creative is to produce an output of novelty and *value*. It requires knowing the rules that can be broken and themselves played with, not just seeing the shadow the rules leave on games already played.

The importance of context, especially context written by human experts, is explained by understanding that well written context *is* the set of assumptions that can allow genuine

conjecture to be produced by a predictive system²⁵. We should be therefore excited that progress will be accelerated by leading experts in various fields having access to AI research partners. And as we’ve discussed, the continued importance of taste, when properly considered as a talent for nudging the system in interesting directions by applying outside assumptions as input rather than merely judging the outputs produced.

For organisations trying to deploy AI, the question of how to work with this transformational technology is ultimately a question of how employees should intervene on the creative process of conjecture and criticism when the majority of their output is being produced by language models. Whether they build this capability in-house, buy application layer solutions, or bring in consultants from the frontier labs or elsewhere, the continued trend has been to push out the horizon over which AI operates. Human judgement is moving father from the fractal complexity of the work, and closer to the final output. My belief is that finding the “level” at which the universal explainer (the employee) should intervene on the creative process of a piece of work will be the fundamental problem facing organisations of the future. Some tasks will be achievable entirely agentically, while some will require variable levels of human intervention to guide the process of conjecture. This will be a hard balance to strike. Writing anything manually is clearly “inefficient” now, but it is hard to assign a value to the thoughts simulated within the brain doing the writing. This holds for any piece of work. There is an unknowable opportunity cost to the thoughts foregone when one imposes judgement only on the output.

We live in a strange time and strange times attract simple stories. The simple story of this era is that you should just trust the trend line. 5 years ago AI couldn’t string together a coherent sentence. Today it can read your emails, book your flights, and make a passable attempt at almost any piece of work you’re still being paid to do. Imagine what it might be able to do in another 5? The story I’ve told here is, on one hand, more complex. It requires believing in a future where AI can seemingly do almost everything a human can do faster and more cheaply and yet it cannot do the one thing we really need. But on the other hand, it is an equally simple story with a trend line much more established in history. That is the story of the project we call Progress, which has always relied on *creation ex nihilo*: creation from nothing. Such creation arises from an ability which all human brains possess; to perceive the causal territory of reality, to ask how things could be different, and to use our inherent “nose for the interesting” to find the conjecture worth exploring next. It is this ability that constitutes the explanation—the short program—of intelligence. We simply need to figure out how it is written.

²⁵SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks (Li et al, 2026): <https://arxiv.org/pdf/2602.12670>