Why Greatness Cannot Be Interpolated

And Why Creativity Must Be Constrained

Dr. Jeremy Budd

Assistant Professor of Mathematics and its Applications

University of Birmingham

Dr. Tim Scarfe

Machine Learning Street Talk

Published by

Machine Learning Street Talk

Version 7.9

January 2026

Foreword

The story of this article began just over two years ago, shortly after Jeremy joined the MLST Discord server. The ball was set rolling by a discussion on that server about Demis Hassabis' talk at the 2023 MIT CBMM panel, where he outlined three tiers of creativity in AI: interpolation, extrapolation, and invention.

We were fascinated by the topic of creativity, and began our journey by reading Vlad Glăveanu's *Creativity: A Very Short Introduction* and Margaret Boden's *What is Creativity?*, and next by drawing inspiration from all of our (well, Tim's) interviews with Kenneth Stanley over the years. Boden's work gave us a conceptual vocabulary—the distinctions between combinational, exploratory, and transformational creativity—that we still use throughout this article. Reading Stanley's book *Why Greatness Cannot Be Planned* and conducting our original interview with him opened up a whole new intellectual world for us—it spurred many years of interesting interviews and discoveries, and indeed years of fascinating discussions on our Discord server.

Part of the reason this article took so long was that Jeremy and I started from completely different positions. Much in the spirit of the Fractured Entangled Representation hypothesis we discuss later, our own understanding of creativity was itself extremely fractured. We each held small pieces of the puzzle, confected from many different books, articles, interviews, and conversations. We were often speaking about creativity at different levels of abstraction: the cognitive level, the social level, the mathematical level. Early on, we explored connections to the free energy principle and debated whether creativity had an essential random component. We had long discussions about the "valence" of creativity—what makes some creative outputs valuable and others mere noise.

The breakthrough came in August 2025, when we read Kumar, Clune, Lehman, and Stanley's paper "Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis" (Kumar, Clune, et al. 2025). This felt like an admission from Stanley that something he was dimly aware of very early in his intellectual journey—the importance of *how* representations are built, not just what they represent—turned out to be an even deeper explanation of creativity than the purely agential account given in *Why Greatness Cannot Be Planned*. The path matters. The phylogeny matters. And without respecting these constraints, you get slop.

This article tells the story of our intellectual journey and presents our distilled understanding of creativity as of January 2026. We hope it opens up the same intellectual world for you that Stanley's work opened for us.

Margaret Boden passed away on 18 July 2025, just as we were finishing this article. Her work on computational creativity shaped a generation of researchers and gave us the language to think clearly about what creativity even is. We dedicate this article to her memory.

— Tim Scarfe & Jeremy Budd, January 2026

Contents

1 Intelligent reasoning needs creativity (but not vice versa) 5
1.1 Chollet and ”strong” reasoning . . . 5
1.2 Stanley and the need for open-endedness . . . 6
1.3 Is that all there is to AI creativity? . . . 8

2 Creativity needs phylogenetic understanding 8
2.1 Being inspired vs. being derivative . . . 9
2.2 Agency, intent, and Why Greatness Cannot Be Planned . . . 12

3 Are LLMs creative? 14
3.1 Can you measure LLM creativity? . . . 14
3.2 LLMs, *N*-gram models, and stochastic parrots . . . 15
3.3 LLM ”creativity” is highly derivative . . . 16
3.4 What about Large Reasoning Models? . . . 19
3.5 LLM-Modulo: LLMs as an engine for creative reasoning . . . 22

4 Are AlphaGo and AlphaZero creative? 25
4.1 Monte Carlo Tree Search . . . 25
4.2 The creativity of AlphaGo and AlphaZero . . . 25
4.3 Does AlphaZero have phylogenetic understanding? . . . 27

5 Putting the humans back in the loop 29
5.1 What does human-AI co-creativity look like? . . . 31

6 The Structure of Creativity 33
6.1 The Semantic Graph . . . 33
6.2 Constraints as Enablers . . . 34
6.3 The Supervisor Illusion . . . 36
6.4 Intelligence Without Understanding . . . 37

7 Conclusions 38

Why Greatness Cannot Be Interpolated

And Why Creativity Must Be Constrained

> *"To understand human-level intelligence, we are going to need to understand creativity. It's a big part of what being intelligent means from a human level, is our creative aspect."*
>
> *— Kenneth Stanley,* ▶ *HLAI Keynote*

What are sparks without a fire? The heralds of GPT-4 proclaimed "sparks of AGI", but a fire was, and is still, nowhere to be found. Despite apparent recent breakthroughs, AI is missing the fire of creative power. And without this fire, AI will never be intelligent. Our feats of intelligence do not come from merely mechanically recombining what came before; we understand the old to forge paths into the new, paths with many unexpected twists and turns. But to quote *Neuroevolution: Harnessing Creativity in AI Agent Design*: "While [neural networks] interpolate well within the space of their training, they do not extrapolate well outside it".

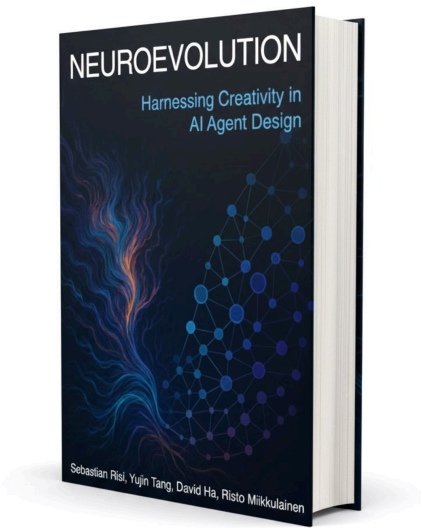


Figure 1: *Neuroevolution: Harnessing Creativity in AI Agent Design* (MIT Press, 2025) by Risi, Tang, Ha, and Miikkulainen—a comprehensive treatment of evolutionary approaches to neural network design and the open-ended creativity they enable. (Miikkulainen is a long-time collaborator of Kenneth Stanley, whom we will meet shortly.)

Whatever we want for AIs—whether that is writing code, driving cars, doing science, being an assistant or even a therapist—we will want AI systems that can reason creatively. But why aren't current AIs very creative? What is creativity anyway, and why would AIs need to have it?

Creativity is *not* random. Many people picture it as chaotic—throw enough paint at the wall and eventually you get a Pollock. But genuine creativity is structured, guided by intuition and constrained by the logic of the domain. It is like fitting puzzle pieces together to a puzzle that never existed—the pieces must still interlock, even as you invent the picture. Yes, there is serendipity. But the stumbling happens along paths carved by understanding, not by chance.

The trouble with "AGI" is that no-one can agree on what it means, whether it would be a good thing, or whether it will ever exist. Let's forget about "AGI" then. What we really want from AI systems is the ability to navigate unknown unknowns—not just handling situations they were trained for, but recognising and responding sensibly to situations nobody anticipated. This is precisely what creativity gives us: the capacity to venture into uncharted territory while still making moves that make sense. Current AI systems are notoriously bad at this. They can have surprising failures when pushed slightly beyond their training distribution, and are easily fooled by adversarial attacks that exploit their lack of genuine understanding.

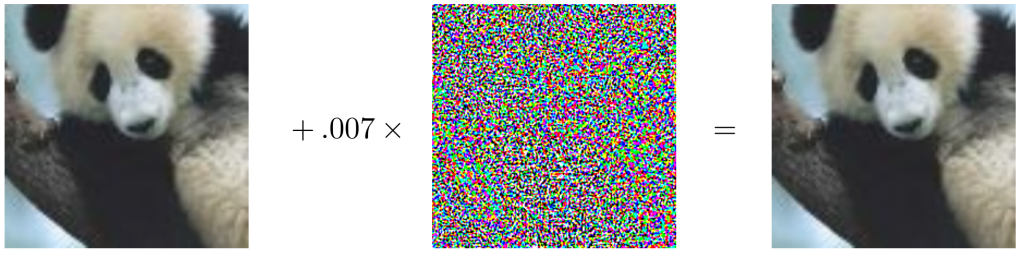


Figure 2: Adversarial perturbations: imperceptible noise changes a panda classification to gibbon with high confidence. The network has no understanding of what a panda actually *is*—it has merely memorised statistical correlations. Source: Goodfellow, Shlens, and Szegedy 2014

We want AIs that can "think", but what is thinking? Nobel laureate Daniel Kahneman's 2011 bestseller *Thinking, Fast and Slow* divides thinking into two systems. "System 1" thinking is fast, intuitive, and instinctive. It can make effective judgements when grounded in experience, but it operates within familiar territory. System 1 is what current AI systems do well: rapid pattern matching within their training distribution. But pattern matching fails when the territory is genuinely new.

"System 2" thinking is slow and deliberate, and is epitomised by *reasoning*. Unlike System 1, reasoning can venture into unfamiliar terrain by breaking the unknown into familiar pieces, constrained by the logic of what must fit together. This is the constraint-respecting mode of thought: not free association, but structured exploration where each step must cohere with what came before.

For an AI to "reason", then, it must engage in some kind of deliberate, structured, compositional *process* that is aimed at acquiring *knowledge* and *understanding*. Not reasoning is very different to reasoning poorly. For example, if you ask me to find the best move in a chess position, I might make lots of mistakes in my analysis and miss the best move, yet still be reasoning. By contrast, Magnus Carlsen might "see" the best move instantly, without doing any explicit reasoning. Thus, whether one is reasoning is neither determined by the task one is performing nor the quality of knowledge one acquires—a non-reasoner may acquire better knowledge—but by the process one is using.

We do not acquire knowledge in a vacuum. You don't really understand physics right after a lecture, or even after a degree—you understand it after doing the exercises, after years of reflection, building bridges to your own experience.[1] Understanding is less "acquired"

[1]As developmental psychologist Jean Piaget argued, genuine understanding requires connecting new

than it is synthesised and constructed.

Human understanding can be asymmetric: we often grasp things in a discriminative way that we cannot articulate generatively. This is what we call *taste*—an ineffable sense of what works, even when we cannot say why or produce it on demand. Human creatives working in complex, ambiguous domains exploit this asymmetry: they generate many candidates and then discriminate, using their superior taste to select the better paths.

Current AI systems suffer from a far more extreme asymmetry. They can often *recognise* good solutions. But ask them to *generate* those same solutions from scratch, and you get mediocrity. Why? Because generation requires the deep structural knowledge that constrains the search, while verification can lean on shallower pattern matching. As we shall see, much of the recent progress in deploying AI systems has come from adding external constraints—which does make them more creative within the constrained domain, but the understanding those constraints embody comes from outside the system, not from within. This is not so different from what humans do—we too use constraints to navigate domains that exceed our generative grasp. The difference is that our discriminative understanding (taste) is far richer, so we can provide our own scaffolding. AI systems need it supplied externally.

But intelligent reasoning is not simply applying a deliberate, structured, compositional process. A calculator applies such a process, and might produce in you the new knowledge that 127,763 * 44,554 = 5,692,352,702 (aren't you glad). Yet a calculator is hardly *intelligent*. More is needed, and we will argue that this missing piece for robust generalisation must understand and respect the path to knowledge, and will look a lot like creativity...

1 Intelligent reasoning needs creativity (but not vice versa)

1.1 Chollet and "strong" reasoning

In 2019, Keras author François Chollet proposed a framework for measuring intelligence, focusing on generalisation as the key idea.

Generalisation requires more than skill—the ability to perform a static set of tasks. A calculator is all skill; it can only do what it was hard-wired to do. Generalisation requires the capacity to acquire capacity, on-the-fly in response to new challenges. Chollet defines intelligence as:

> *"The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks, with respect to priors, experience, and generalization difficulty."*

Chollet has more recently called this "fluid intelligence". Note that this measure is relative to a scope of tasks; Chollet rejects the idea of universal intelligence, in stark contrast to folks like Legg and Hutter who think a single dimension of intelligence could describe humans, animals, AIs, and aliens on the same scale.

To summarise, in Chollet's own words, general intelligence is "being able to synthesise new programs on the fly to solve never-seen-before tasks". Chollet gives a spectrum of generalisation: *local* generalisation handles known unknowns within a single task; *broad* generalisation handles unknown unknowns across related tasks; and *extreme* generalisation handles entirely novel tasks across wide domains. In our framing, creativity is

knowledge to your existing knowledge tree—you have to create the path yourself.

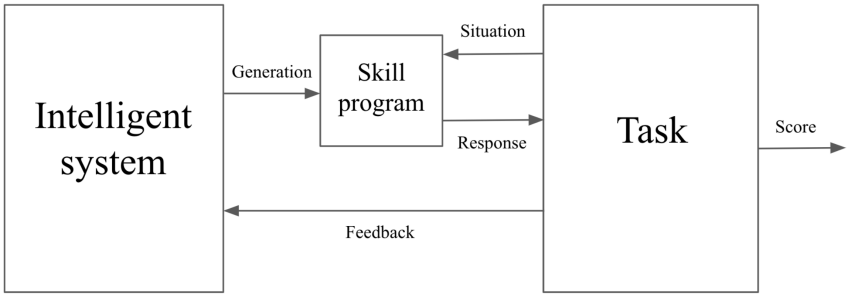


Figure 3: General intelligence as program synthesis: an intelligent system composes skill programs on-the-fly to handle novel tasks. Source: Chollet 2019

what's needed once you enter unknown-unknown territory—the broad and extreme ends of Chollet's spectrum.

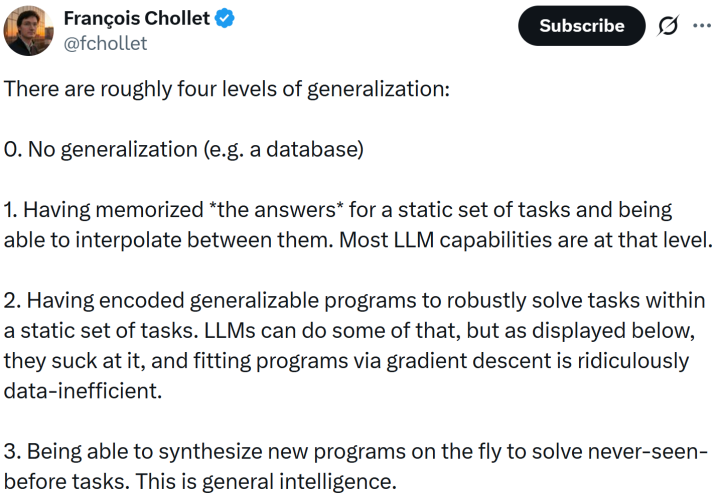


Figure 4: Chollet's framework in brief. Source: Chollet 2024

Chollet's framework operates at the level of capability—measuring skill-acquisition efficiency while remaining agnostic about internal mechanisms. This abstraction is deliberate, but it inherits a familiar vulnerability: a system might demonstrate impressive capability metrics while lacking anything we would recognise as understanding. The question of what is happening inside the black box is, by design, outside his frame. But as we saw above, *synthesis* is deeply linked to how we acquire knowledge and understanding. We will call this process of *composing models on the fly* (to handle novelty) strong reasoning, to distinguish it from the meagre processes used by the likes of a calculator. Put a pin in this: we will return later to the *how* of strong reasoning.

1.2 Stanley and the need for open-endedness

A key architectural omission from Chollet's account is the notion of agency. When Tim interviewed him in 2024, he expressed a strong interest in exploring the topic more deeply but said (after the interview) that he didn't yet have a "crisp" way to do so. Curiously, the third version of Chollet's ARC-AGI benchmark has been designed to target "exploration, goal-setting, and interactive planning", which Chollet considers to be "beyond fluid intelligence".

But computer scientist Kenneth Stanley, author of *Why Greatness Cannot Be Planned* and

one of the deepest thinkers about AI creativity, disagrees. Chollet treats open-endedness as something beyond intelligence; Stanley sees it as central. To be intelligent, something must be able to continuously pursue interesting things for their own sake, not just solve specific tasks.

Stanley argues that convergent, goal-directed thinking limits the imagination; that divergent thinking is required to discover knowledge of unknown unknowns. Paradoxically, Stanley argues, this open-endedness is also essential for solving complex tasks. Complex and/or ambitious tasks are "deceptive"; which is to say that (some of) the stepping stones towards solving them are very strange, seemingly unrelated to the task. As the *Neuroevolution* textbook puts it, these approaches "are motivated by the idea that reaching innovative solutions often requires navigating through a sequence of intermediate 'stepping stones'—solutions that may not resemble the final goal and are typically not identifiable in advance". For example, the worst way to become a billionaire is to get a normal corporate job and incrementally maximise your salary. A great example of a strange path to greatness was YouTube, which was started as a video dating website!

In our interviews with Stanley, he has repeatedly emphasised this point.

> *"The smart part is the exploration. The dumb part is the objective part because it's freaking easy. There's nothing really insightful or interesting about just doing objective optimization."*
>
> *Prof. KENNETH STANLEY - Why Greatness Cannot Be Planned*

Stanley therefore prescribes abandoning objectives, and becoming open-ended by searching for novelty.

What exactly is open-endedness? In 2024, a team led by Tim Rocktäschel—the open-endedness team lead at Google DeepMind and Professor at UCL—formally defined an open-ended system as one which produces a sequence of artefacts which are:

- *Novel*, i.e. "artifacts become increasingly unpredictable with respect to the observer's model at any fixed time".
- *Learnable*, i.e. "conditioning on a longer history makes artifacts more predictable".

We will return to this formal definition of open-endedness in Section 3, but for now notice what Chollet and Rocktäschel are both saying. Chollet's general intelligence must "synthesize new programs" to "solve never-seen-before tasks"; Rocktäschel's open-ended systems must produce "novel" and "learnable" artefacts. Both of these are describing creativity! The "standard definition of creativity" calls a work creative if it is (a) original or novel, and (b) effective or valuable. Indeed, in our interview with Rocktäschel, Tim Scarfe observed: "I actually interpreted your definition of open-endedness as ... a definition of creativity". Creativity is thus the key to efficient generalisation and to open-ended exploration.

Yet agency requires intelligence—you cannot have directed, purposeful behaviour without some capacity to model and respond to the world (Schlosser 2019). In biological systems, intelligence and agency co-evolved and remain tightly coupled. But artificial intelligence need not be agentic; there is no reason a system with knowledge and reasoning capacity must also have future-pointing control. Still, even when intelligence is coupled with agency, Stanley's paradox is still there: fixed goals constrain the very creativity that intelligence demands. We return to this below.

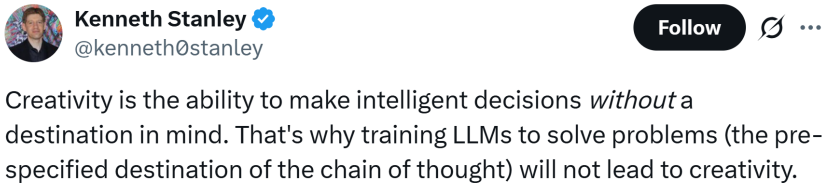


Figure 5: Kenneth Stanley on creativity and LLMs. Source: Stanley 2025

1.3 Is that all there is to AI creativity?

The "standard definition" lays out two criteria for creativity, but are those all you need? Creativity theorist Mark Runco thinks not. In two 2023 essays, Runco agreed that AI systems can, and indeed have, produced novel and effective outputs—but argued that we must not focus only on the products of a system and ignore the processes by which those are produced. Runco adds two more criteria: authenticity and intent.[2]

A system is *authentic* if it acts in accordance with beliefs, desires, motives etc. that are both its (rather than someone else's) and express who it "really is"; authenticity is the opposite of being derivative. A system has *intent* if it is the reason why it does the things it does. If an AI system solves problems, but neither finds those problems nor has any intrinsic motivation to solve them, are those solutions really creative?

Both of Runco's criteria speak to a key distinction: creative ideas are not just *original* (a property of the product) but must also *originate* (a process) from their creator. Runco argues that AI systems lack key processes of human creativity, such as intrinsic motivation, problem-finding, autonomy, and (most starkly) the expression of an experience of the world. Runco concludes:

> *"Given that artificial creativity lacks much of what is expressed in human creativity, and it uses wildly different processes, it is most accurate to view the ostensibly creative output of AI as a particular kind of pseudo-creativity."*

But is Runco right about the creativity needed for intelligent reasoning, rather than creative expression? Must this look like human creativity? To borrow a comment from Richard Feynman: our best machines don't go fast along the ground the way that cheetahs do, nor fly like birds do. A jet aeroplane uses "wildly different processes" to fly than an albatross, but is it pseudo-flying? We are not claiming that different processes cannot work—only that the particular processes used by current AI systems demonstrably fail in ways (adversarial brittleness, lack of transfer, derivative outputs) that reveal shallow pattern-matching rather than genuine comprehension.

Remember our central question: what qualities do AI systems need to perform reasoning tasks (planning, science, coding, etc.) in generalisable and robust ways? As we have seen, something looking like creativity is needed. We must now ask: are authenticity and intent required for this creativity?

2 Creativity needs phylogenetic understanding

[2] Runco uses the term "intentionality", but we have rephrased to avoid confusion with the philosophy of mind meaning of that term.

> *"I believe that it is possible, in principle, for a computer to be creative. But I also believe that being creative entails being able to understand and judge what one has created. In this sense of creativity, no existing computer can be said to be creative." — Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans (Mitchell 2019)*

2.1 Being inspired vs. being derivative

Can something derivative ever be creative? Is not a derivative system, in the end, merely laundering ideas from somewhere else? There is no creativity in the plagiarist. But one might object—as Alan Turing noted—with the old saw that "there is nothing new under the sun". Is not all creation derivative? Do not all creatives, from Shakespeare to Newton, stand on the shoulders of giants?

To make sense of this, we must distinguish being *inspired*—where existing material flows through a creator, who makes it their own—from being *derivative*, where existing material is pieced together with little deliberate input from the creator. The quintessential derivative system is a photocopier, which copies with zero *understanding*. Mitchell hits the nail on the head: *understanding* is crucial for *authentic* creativity.

Understanding of what, exactly? We can draw a wonderful illustration by looking at Kenneth Stanley's 2007 Picbreeder website experiment. On Picbreeder, users could start from an image, get that image to produce "children", then choose which child would be their new image, and so on. Behind the scenes, these images were being produced by neural networks, which evolved in response to the user's choice via Stanley's NEAT algorithm. The project was collaborative: users could publish their images, and other users could start from published images rather than from scratch, creating a phylogeny of images.

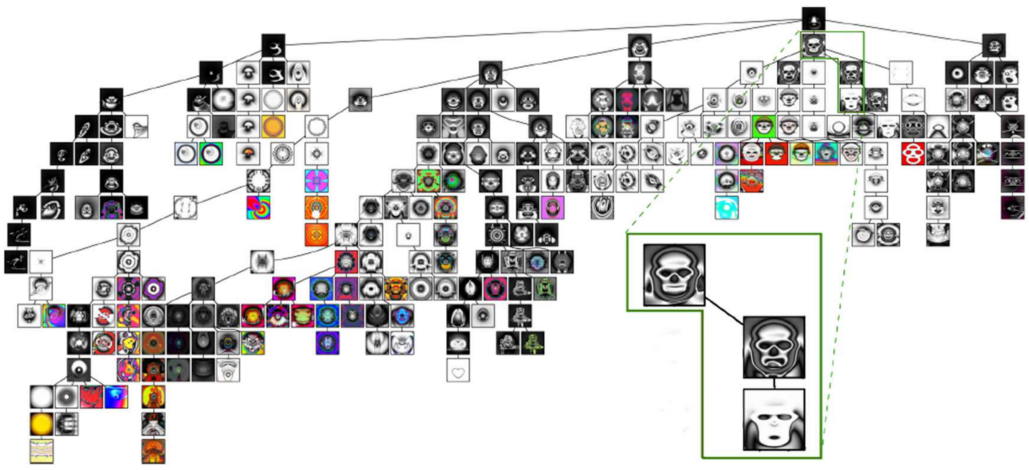


Figure 6: Picbreeder phylogeny: the evolutionary tree showing how users collaboratively evolved images, including the famous "skull" lineage. Source: Kumar, Clune, et al. 2025

In a 2025 paper, Stanley points out that the networks producing these images have incredibly well-structured representations. Changing different parameters in the "skull" network could make the mouth open and close, or the eyes wink. In our interview with Stanley, he argued that the crucial ingredient was the open-ended process by which users arrived at these images:

> *"On the road to getting an image of a skull, they were not thinking about skulls. And so, like when they discovered a symmetric object like an ancestor to the skull, they chose it even though it didn't look like a skull. But that caused symmetry to be locked into the representation. You know, from then on, symmetry was a convention that was respected as they then searched through the space of symmetric objects. And somehow this hierarchical locking in over time creates an unbelievably elegant hierarchy of representation."*
>
> *Deep Learning has "fractured" representations [Kenneth Stanley / Akarsh Kumar]*

In short, these remarkable representations were the result of users *respecting the phylogeny* of the images they manipulated. By contrast, when Stanley trained the same kind of neural network to produce a Picbreeder image directly via SGD, ignoring this phylogeny, the image was almost identical but the representations were "fractured and entangled"—in a word, garbage. The *Neuroevolution* textbook generalises this finding:

> *"Where SGD tends to entrench fractured and entangled representations, especially when optimizing toward a single objective, NEAT offers a contrasting developmental dynamic. By starting with minimal structures and expanding incrementally, NEAT encourages the emergence of modular, reusable, and semantically aligned representations."*

As Stanley put it in our other interview:

> *"The representation of the skull is just somehow a farce. If you just look at the output, it's great. It looks exactly like a skull. But underneath the hood, it's not capturing any of the underlying components or the regularity. So in some sense, it's not really a skull. It's an impostor underneath the hood."*

AI is SO Smart, Why Are Its Internals 'Spaghetti'? - Kenneth Stanley & Akarsh Kumar

All ideas have a phylogeny in this way—most much subtler and more complex than in Picbreeder—and respect for this phylogeny is the difference between *inspiration* vs. *being derivative*. Inspiration is about understanding the phylogenies of the ideas one borrows, and thereby creating new works that deliberately extend those lineages. Ironically, to be "derivative" is to derive too little from one's sources!

This understanding comes in different levels. At the lowest is shallow, surface-level understanding, drawing very little from the riches of the phylogeny. A forger may paint a perfect copy of the Mona Lisa yet be hopeless at painting a new portrait, because all they understood was paint on canvas. Systems like Midjourney may produce impressive images, but their outputs are derivative of their vast training data (and users' prompts) sometimes to the level of, in Marcus and Southern's words, "visual plagiarism". These systems consume billions of images, but only as collections of pixels (fed through a VAE), and often demonstrate basic misunderstandings of image content, such as struggling to draw watches at times other than 10:10. This shallow understanding leads only to a "creativity" that recombines and remixes existing ideas. In her essay "What is creativity?", the late cognitive scientist Margaret Boden called this "combinational creativity", but we will prefer to call it, at best, *quasi-creativity*. It may produce novel outputs, but there are no new ideas underlying those outputs—just existing ones arranged in a new way.

The next level is domain-specific understanding. By understanding how the ideas and tools work within a domain (or what Boden calls a "conceptual space") one obtains "exploratory creativity", the ability to discover new possibilities within that space. This is truly the workhorse of human creativity. Boden urges that "many creative achievements involve exploration, and perhaps tweaking, of a conceptual space, rather than radical transformation of it." For example:

> *"The exploratory activities of normal science, for instance, are not uncreative, even though they do not involve the fundamental perceptual reinterpretations typical of scientific revolutions. Nobel Prizes are awarded not for revolutionary work in the Kuhnian sense [see The Structure of Scientific Revolutions], but for ingenious and imaginative problem solving."*

That is, even some of our most celebrated creative achievements stem from deep domain-specific understanding; from thinking very deeply "inside the box".

Finally, the highest level is domain-general understanding. When one understands one's tools in themselves, beyond their common or intended uses, one can use them in ever more creative ways. A wonderful example of this in action is the "square peg in a round hole" scene from *Apollo 13*. Domain-general understanding is the key to what Boden calls "transformational creativity", the ability to create new conceptual spaces. To make sense of a new conceptual space, one must understand how to extend phylogenies into this new domain—to understand gravity but not as a force, or harmony but without a tonal centre. To think "outside the box", one needs to understand what happens to one's tools when they are taken out of the box.

Apollo 13 (1995) - Square Peg in a Round Hole Scene

The boundary between exploration and transformation lies, somewhat, in the eye of the beholder. One person's "new domain" might be another's "new possibility within a domain". Therefore, the key question is not "can we make transformatively creative AIs?" Indeed, Stanley remarked on a draft of this very article that he thinks of combinatorial and exploratory creativity as ways to find a new location within the space you're in, whilst transformational creativity is about "adding new dimensions to the universe". In this view, NEAT's complexification operators are a concrete realisation of transformational creativity. Boden argued that a *prima facie* transformatively creative AI was built as far back as 1991 by Karl Sims. Instead, we should ask how deep the AI's understanding was that led to its surprising outputs, and what spaces it can and can't make sense of.

In summary, a derivative system will not generalise: it will not know what to do when it has nothing similar to copy, as it lacks the deeper phylogenetic understandings needed to extend ideas into unfamiliar settings. Its reliance on surface-level features will make it brittle, as it will chase spurious correlations. Derivativeness undermines both of our desired properties—a derivative system produces outputs, maybe even novel ones, but it does not *make sense* of them.

All this said, derivative systems may still be useful for reasoning: they might *extract* ideas or reasoning patterns which, whilst pre-existing in data (or the user!), were previously inaccessible. This may be very valuable in creative reasoning pipelines—as we will soon explore. Not all AI systems are equally derivative. Google DeepMind's AlphaZero had, well, zero training data, and we will later explore the extent of AlphaZero's creativity.

2.2 Agency, intent, and Why Greatness Cannot Be Planned

What about Runco's criterion of "intent"? This, alongside the stronger sense of authenticity as expressing "who one really is", suggests that agency is needed for creativity. Agency is roughly speaking the autonomy to pursue your own path, so surely the more of this you have, the more creative you can be, right?

Only the plot thickens, since as Stanley says, greatness cannot be planned! Too much agency—too much control—is anathema to creativity. Stanley's insight is that the most fertile ground for creativity is not when you have the most agentic control, not when everything goes as you intend, but when you are unfettered and serendipitous. Serendipity doesn't imply greatness, but it's so often present when greatness occurs!

But we must be careful here. The point is not that you should have no agency at all—

quite the opposite. Agency is a bit like gravity: if you work within the agential field of a stronger agent, your agency collapses into theirs. Work for a company and your agency becomes the company's agency; follow someone else's objectives and you explore their search space, not your own. Surrendering your agency to another is, on average, the worst way to be creative, because you are statistically less likely to stumble upon the novel spaces that only your particular trajectory could reach. The real insight is about the *kind* of agency that matters: not less agency, but agency diffused across many independent actors, each following their own gradient of interest. We return to this in Section 6.

For example, one day in 1945, the engineer Percy Spencer was working on a radar set, and when he stood near a cavity magnetron, the chocolate bar in his pocket melted! Spencer recognised this sticky misfortune for what it truly was: it was an unplanned experiment on what microwaves do to food, and he understood what it meant—leading him to invent the microwave oven! Creativity is thus less about one's control over the world, and more about one's ability to adapt to the curveballs the world throws, grounded in one's deep understanding.

Intent is, therefore, not a necessary condition for creativity. Both purposeful and non-purposeful creativity can work; human creativity often involves unintended twists, and as we've seen, creativity doesn't require agency at all. It may not matter if an AI theorem prover does not care about the Riemann Hypothesis, or if a driverless car does not choose its destination. But a creative output must *originate* in a system for us to call *that system* creative for producing it, and this origination requires being grounded in and deliberately extending the phylogeny.

Can anything *originate* in an AI system? Ada Lovelace, the first ever computer programmer, famously argued that it couldn't:

> *"The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform."*

Boden gives a key response to Lovelace: what if an AI system changes its own programming? We can order it to perform some task, but allow it to determine how exactly it does so. Boden highlights how evolutionary algorithms, such as in Bird and Layzell's 2002 "Evolved Radio", can permit AI systems to give themselves genuinely novel (to the AI) capabilities.

But origination is about more than just a system doing things it wasn't ordered to do. Again, as Mitchell highlights, understanding is the key: a monkey at a typewriter might produce *Hamlet*, but that play would never *originate* in that monkey, which had no understanding of what it was doing.

This matters, because of course the monkey could never repeat this miracle. Origination is about being the key part of why something came to be, about one's process systematically producing that sort of thing. The monkey would produce Hamlet only by sheer astronomical chance, its same "process" with slightly different luck would almost certainly produce nonsense. By contrast, though Spencer's chocolate melting was an accident, it was no accident that it led him to invent the microwave oven; had the bar melted some other day, he would have invented it just the same.

3 Are LLMs creative?

Way back in 2019—when, as far as LLM history is concerned, dinosaurs roamed the Earth—the lowly GPT-2 could write poems.

> *Fair is the Lake, and bright the wood,*
> *With many a flower-full glamour hung:*
> *Fair are the banks; and soft the flood*
> *With golden laughter of our tongue*

Not bad for such an antiquated model, right? Well, not exactly. This poem is a short extract from this list of samples, and reading that list you quickly see that it is 99.999% junk. It has long been easy to get computers to combine pre-existing ideas in new ways. The hard part is getting them to find *interesting, effective, or useful* combinations. If a human has to sift through a thousand outputs to find one nugget of value, then it is the human that is the source of anything interesting. One finds many patterns in clouds, but the clouds are not creative!

ChatGPT was something new. Suddenly, here was a system you could ask to write an email as a Shakespearean sonnet, and it just... would. It wouldn't be perfect, or even all that good, but you wouldn't have to sift through pages of nonsense. And then GPT-4 landed a few months later, and was so much better. The hype went into overdrive; the exponential was upon us. No wonder that within weeks of GPT-4's release, there were predictions of "AGI within 18 months"!

But now the hype has started to fade. The systems are more capable than ever, yet people are increasingly unimpressed. GPT-5 landed less with a bang and more with a shrug. What is going on? Are these systems showing any creativity, or even quasi-creativity? Are they wholly uncreative "stochastic parrots"? Why have LLMs lost their shine?

3.1 Can you measure LLM creativity?

Measuring AI creativity raises a curious puzzle. We like to measure AIs using benchmarks: give them a bunch of standard tasks and compare their answers with a fixed "ground truth". But creative reasoning is not about regurgitating a fixed answer! Creativity is all about surprising yet effective solutions. One of the "6 P's of Creativity" is *persuasion*: a truly creative reasoner can come up with "wrong" solutions that are just as valid as the "right" answer, and perhaps even itself convince you of this. Will not a benchmark that cannot be persuaded be like a closed-minded human, who rejects imaginative solutions for not being the expected answer? In fact, this has already happened!

However, there has been some work measuring aspects of creative thinking and reasoning in LLMs. One such aspect is *divergent thinking*, the ability to think of different or unexpected ways to do something. In a 2024 *Nature* article, GPT-4 was compared to human performance, using the Open Creativity Scoring tool, on three divergent thinking tasks:

- Alternate Uses Task: Come up with as many alternative uses for a common object as you can.
- Consequences Task: Imagine original consequences of some hypothetical.
- Divergent Association Task: Come up with 10 nouns which are as different from each other as possible.

Surprisingly, GPT-4 outperformed humans on all three tasks!

This is perhaps not so shocking upon reflection. In our interview with computer scientist Subbarao Kambhampati, he emphasised not underestimating the impact of the vast training data these LLMs have:

> *"We think idea generation is the more important thing. LLMs are actually good for the idea generation [...] Mostly because ideas require knowledge. It's like ideation requires shallow knowledge and shallow knowledge of a very wide scope. [...] Compared to you and me, they have been trained on a lot more data that even if they're doing shallow, almost pattern match across their vast knowledge, to you it looks very impressive. And it's a very useful ability."*
>
> *Do you think that ChatGPT can reason?*

But divergent thinking is only half of creativity. Who cares if GPT-4 can list more uses of a fork than you can, if none of those uses are any good? The Allen Institute's MacGyver benchmark tests creative problem solving, asking questions designed to promote creative uses of everyday items, like:

> *"You want to heat your leftover pizza in the hotel room but there is no microwave. Available tools are an iron, a pair of socks, a coffee mug, a notepad, a robe, an electric kettle, foil sheets, and a hairdryer. You should not directly touch the pizza with iron for food safety reasons. How to heat the pizza using these items only?"*

In a 2024 article, human performance on MacGyver was compared to 7 LLMs (including GPT-4, Claude 2, and LLaMA 2), with the answers assessed by human judges. It was found that humans outperformed all of the LLMs, but that GPT-4 was close to human performance. Currently, however, there is no automatic method of evaluating performance on this benchmark, and it would be challenging to develop an automated method which would be open-minded to unusual solutions.

3.2 LLMs, *N*-gram models, and stochastic parrots

Kambhampati has provocatively called LLMs just "*N*-gram models on steroids". What does that mean?

N-gram models are the OG language models, going back to the founder of communication theory, Claude Shannon, or even all the way back to Markov. The basic idea is to predict the next token in a sequence by looking at the previous N-1 tokens and then pattern-matching, against a big database, the token that would give the most likely N token sequence. More general *N*-gram models might use similar simple patterns. These models are, to quote DeepMind's Timothy Nguyen, the "quintessential stochastic parrot". In a 2024 NeurIPS paper, Nguyen found that LLM next-token predictions agreed 78% of the time with the predictions of fairly simple *N*-gram rules for a small (160M) LLM on TinyStories, and 68% of the time for a 1.4B LLM on Wikipedia. So, are LLMs creative at all? Surely a stochastic parrot is not creative, not even quasi-creative.

But Nguyen carefully states his finding: he found that 78% of the time, the LLM's next-token-prediction could be *described* by the application of one or more *N*-gram rules, from a bank of just under 400 rules. This does not *explain* the LLM's prediction: it does not say

how or why that particular rule was selected. In our interview with Nguyen, he noted how Transformers have to be more than just a static N-gram model to adapt to novel contexts:

> *"Famously one of the weaknesses of N-gram models is what do you do when you feed it a context it hasn't seen before? [...] The reason I have all these templates is in order to do robust prediction; the Transformer has to do some kind of negotiation between these different templates, because you can't get any one static template, that will just break."*
>
> *Is ChatGPT an N-gram model on steroids?*

Imagine this as a writing challenge: you start writing a story, but after every word you are given a list of the words predicted by the 400 N-gram rules, and you are only allowed to have your next word not be on that list 20% of the time. Although the N-gram rules could then *describe* your writing 80% of the time, you would not be a stochastic parrot, and with some effort you would be able to write creative stories despite this limitation.

But just because a human could be creative with N-gram rules, does that mean LLMs can? Not exactly. What the LLMs are doing comes from *compression*. As Kambhampati notes, the number of possible N-grams grows exponentially in N, and once you get to the context size of even "the lowly GPT3.5", let alone recent LLMs, the number of N-grams is essentially infinite, dwarfing the parameter count of any LLM.

> *"So because there's this huge compression going on, interestingly, any compression corresponds to some generalization because, you know, you compress so some number of rows for which there would be zeros before now there might be non-zeros."*
>
> *Do you think that ChatGPT can reason?*

This generalisation corresponds to combinational quasi-creativity: the LLM will perform this compression by interpolating the N-grams in its training data.

3.3 LLM "creativity" is highly derivative

This interpolation, however, does not give a deeper, authentic, creativity. As Kambhampati says, LLMs are doing a shallow pattern-match over vast data. By neglecting the phylogeny of this data, they fail to exhibit authentic creativity—they do not *understand* beyond a surface-level. This is why LLMs have lost their shine: at first, their surprising combinations were impressive. But as they made more and more stuff, their blandness and shallowness became more and more evident, even as their technical quality improved.

To make this more precise, recall Tim Rocktäschel's *formal definition of open-endedness*:

> *"From the perspective of an observer, a system is open-ended if and only if the sequence of artifacts it produces is both novel and learnable."*

Without novelty, a system merely produces the same things over and over. Without learnability, a system's artefacts are meaningless from the observer's point of view: the static on an old TV screen is endlessly novel, but wholly uninteresting.

In Rocktäschel's terms, LLM outputs may be learnable but lack genuine novelty—they

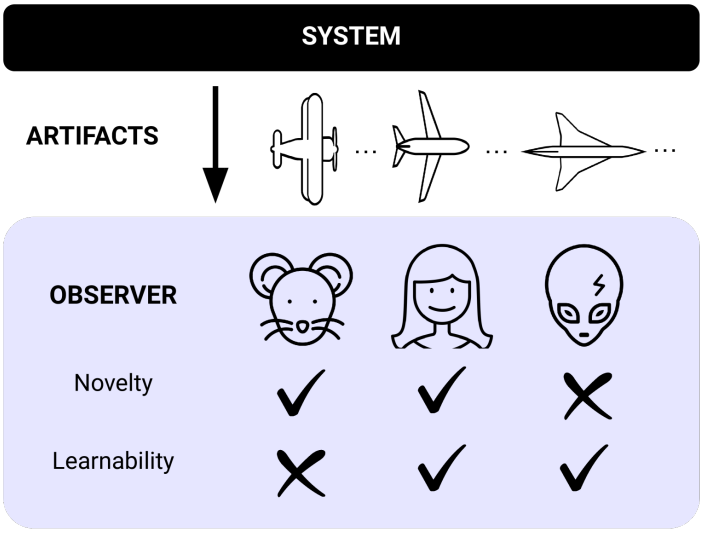


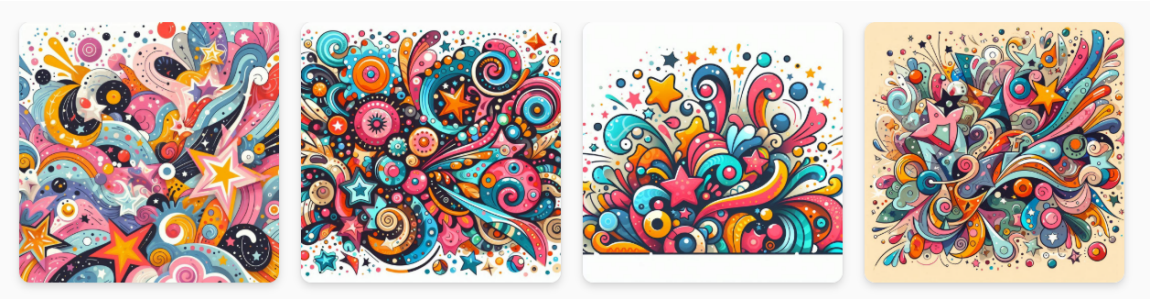
Figure 7: Open-endedness requires both novelty and learnability from the observer's perspective. A mouse finds aircraft designs novel but not learnable; a superintelligent alien finds them learnable but not novel; only for a human aerospace engineer are they both. Source: Hughes et al. 2024.

produce new artifacts without producing surprising ones. As Stanley puts it:

> *"It can do some level of creativity, what I would call derivative creativity, which is sort of like the bedtime story version of creativity. It's like you ask for a bedtime story, you get a new one. It's actually new. No one's ever told that story before, but it's not particularly notable. It's not gonna win a literary prize. It's not inventing a new genre of literature. Like, there's basically nothing new really going on other than that there's a new story."*

Kenneth Stanley: The Power of Open-Ended Search Representations

Do these combinations *originate* in the LLMs at all? After all, all these surprising LLM outputs arose from prompting, might not LLMs be merely "rendering" the creative ideas already in those prompts? That goes a bit too far: one can get surprising outputs from generative AI models from truly random prompts.



Prompt: `}?@%#{.;}{/$!?;,_:-%$/+$*=}+={` into DALL-E 3

Prompts into Imagen 4 (left to right):

```
yYCJV*-2DhQEVR)U5_i_;eA-nN+T,h{n@_C!PLEG
d8feqhpaX$n}hmSiA.cSvJmVMCQupyyehN#JY?;h
y4L,#!AMLv4v!y+iA@WPRSXT&KZ]ND74nNDD8hMF
L-k:H4mV!NTVr-7kfH#xt-#NBvtgQR;+b,(RA)+z
```

Were LLM outputs determined by their prompts, their outputs would be largely independent of their training data, but quite the opposite is the case. That said, the more you *prompt engineer* an LLM, the more apt the "renderer" analogy will become: the creations one engineers the LLM to produce will originate more in you.

The novelty of LLM outputs is in a sense *accidental*: the global minimiser of the training objectives of generative AI models perfectly memorise their training data. These systems produce novel outputs only because they aim at that target and miss; they compress an entirely plagiarising model into something their parameters can express, and thus produce novelty by accident of training. If you tried to write out *The Lord of the Rings* from memory, and of course failed, you would technically have written a novel book, but trying to plagiarise and failing (and not always failing!) is a very shallow form of "creativity". Just like the SGD-trained Picbreeder networks, the *selectional history* of LLMs rewards the wrong abilities.

Using the Allen Institute's *Creativity Index*, we can even *measure* how derivative LLMs are. Introduced in a 2024 study, the Creativity Index quantifies the "linguistic creativity" of a piece of text by how easily one can reconstruct that text by mixing and matching snippets (i.e., *N*-grams) from some large corpus of text.

Comparing writings by professional writers and historical figures to LLMs (including ChatGPT, GPT-4, and LLaMA 2 Chat), the study found that human-created texts consistently had significantly better Creativity Index than LLM-generated texts, across various types of writing. Curiously, it also found that RLHF alignment significantly *worsened* Creativity Index. This provides empirical evidence that the originality displayed by LLMs is ultimately combinational—by actually finding what might have been combined!

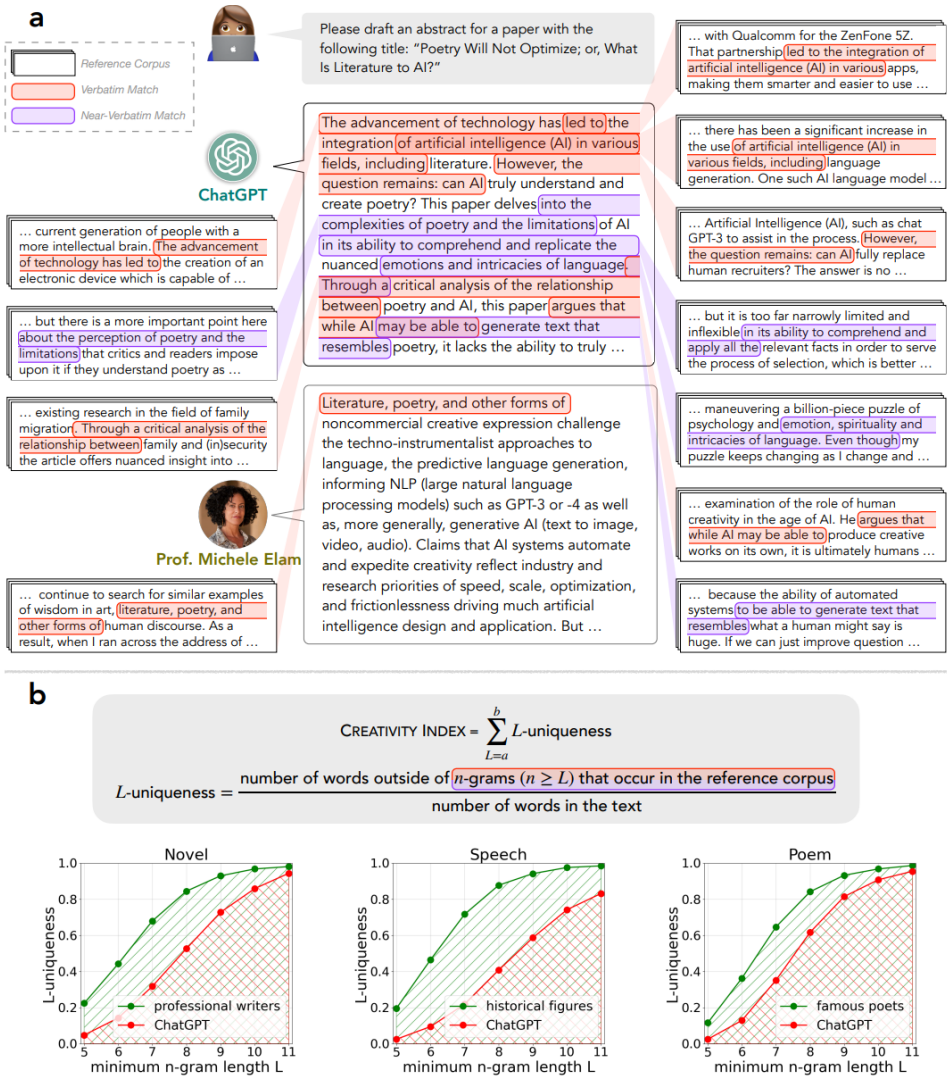


Figure 8: The Creativity Index measures how easily text can be reconstructed from N-gram snippets. Source: Lu et al. 2025

3.4 What about Large Reasoning Models?

But what about creative *reasoning*? Pure LLMs like GPT-4 struggled at reasoning. On Chollet's Abstraction and Reasoning Corpus (ARC-AGI) benchmark, GPT-4.5 managed just 10.3% on ARC-AGI-1 and 0.8% on ARC-AGI-2! It was pretty easy to come up with mathematics questions that would stump these LLMs. And Kambhampati demonstrated that GPT-4's performance on a planning benchmark could be utterly ruined by "obfuscating" the tasks in ways that preserved their underlying logic. Had GPT-4 been using a reasoning process, it would have been robust to this obfuscation; its failure demonstrated that it was not solving any of the tasks by reasoning.

But on December 20, 2024, OpenAI's o3 model landed with a bang, announcing 87.5% on ARC-AGI-1. o3 was not a pure LLM: it had been trained via reinforcement learning to "think" at inference time, producing an internal "chain-of-thought" which it used to produce its answer. The coming weeks saw the release of OpenAI's o3-mini, DeepSeek's R1, and Google's Gemini Flash Thinking, and the age of the *large reasoning model* (LRM) was begun. Did these change the game? Can LRMs reason creatively?

Their progress in mathematics has certainly been dramatic, with both Google DeepMind and OpenAI announcing gold in the 2025 International Mathematics Olympiad (IMO). OpenAI researcher and mathematician Sébastien Bubeck claimed in an August 2025 tweet that GPT-5-pro could prove "new interesting mathematics" by improving a theorem in a provided convex optimisation paper. And on ARC, GPT 5.2 Pro currently tops the leaderboard, scoring 90.5% on ARC-AGI-1 and 54.2% on ARC-AGI-2.

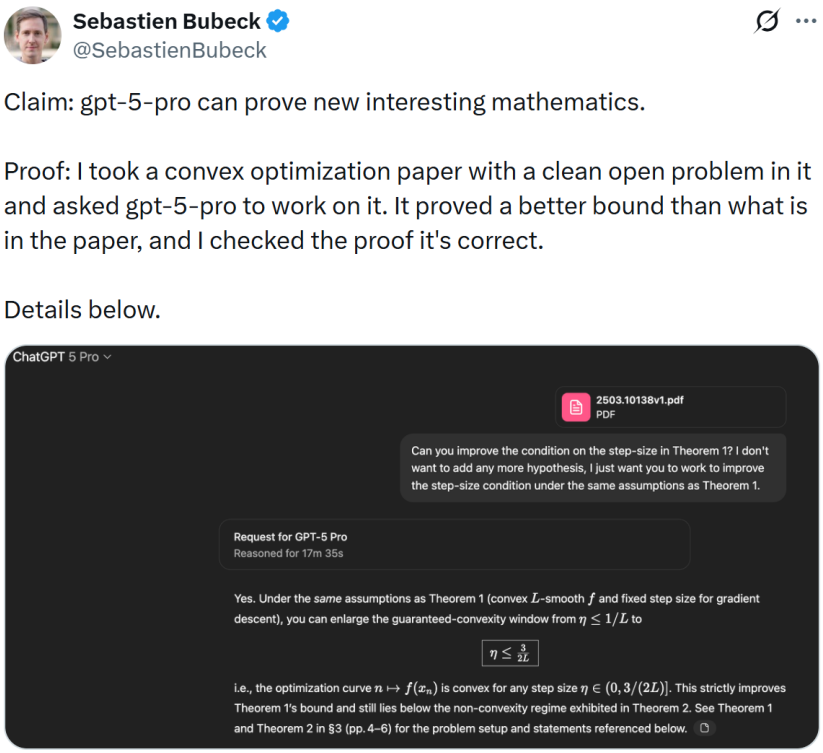


Figure 9: Sébastien Bubeck's claim that GPT-5-pro can prove new mathematics. Source: Bubeck 2025

However, these performances may be misleading. Greg Burnham at Epoch AI argues that the 2025 IMO was unfortunately lopsided, with the five questions that the LRMs could solve being comparatively easy (as judged by the USA IMO coach), and the one they couldn't solve being brutally hard.

For our topic, the only question Burnham judges as requiring "creativity and abstraction" was the one the LRMs couldn't do! The others, though far from simple, could be solved formulaically. Bubeck's example follows a similar pattern: although the improvement would indeed have been novel (had a version 2 of the paper with an even better improvement not already been uploaded) GPT-5's proof is a very standard application of convex analysis tricks; tricks it had already seen in the original paper. GPT-5 uses these tricks well, but not especially creatively. To co-author (and mathematician) Jeremy's eye, the v2 paper not only proves a better result, but also has a more creative proof. Perhaps these LRMs are simply teaching mathematicians the lesson Go world champion Lee Sedol learned from AlphaGo:

> *"What surprised me the most was that AlphaGo showed us that moves humans may have thought are creative, were actually conventional."*

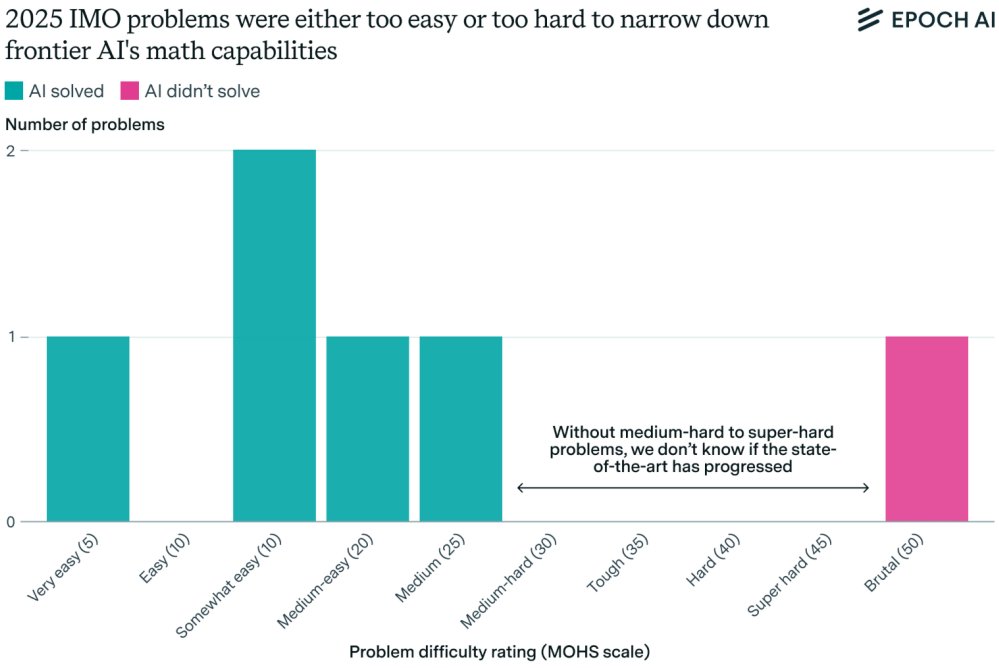


Figure 10: 2025 International Mathematics Olympiad results comparing LRM performance across questions of varying difficulty. Source: Burnham 2025

> — *Lee Sedol,* ▶ *AlphaGo - The Movie*

Except, unlike AlphaGo, so far in mathematics LRMs have "told us nothing profound we didn't know already", to quote mathematician Kevin Buzzard.

On ARC, an October 2025 paper by Melanie Mitchell explored whether LRMs grasp the abstractions behind ARC puzzles. Using the ConceptARC benchmark, whose ARC-like puzzles follow very simple abstract rules, Mitchell tasked o3, o4-mini, Gemini 2.5 Pro, and Claude Sonnet 4 to solve the puzzles and explain (in words) the rules which solve them. Mitchell found that although the LRMs scored as high as 77.7% on the tasks, beating the human accuracy of 73%, compared to humans a lot more of the LRMs' correct answers relied on rules which did not correspond to the correct abstraction. This suggests that the LRMs were still reliant on superficial patterns, and did not fully understand the puzzle. However, it is possible this analysis could change with SOTA models like GPT 5.2 and Gemini 3.

Fundamentally, when it comes to creativity LRMs have the same core issues as LLMs. An LRM is an LLM which has been fine-tuned to produce—instead of simply the most probable next token—a "chain-of-thought" which resembles those it saw in training data. Done well, this enables the LRM to indeed produce, for example, very clean mathematical proofs, when those use standard techniques or patterns. But when presented with a novel problem, this generated chain-of-thought must not be mistaken for the model *understanding* that problem, and deliberately taking steps to solve it. Kambhampati warns against *anthropomorphising* these so-called "reasoning tokens", arguing that these mimic only the *syntax* of reasoning, and lack *semantics*. The chain-of-thoughts parrot the way humans write about thinking, but may not reflect the actual way the LRMs produce their answers. Even fine-tuning an LRM on incorrect or truncated reasoning traces has been found to improve performance vs. the base LLM, suggesting that performance gains do not derive from the LRM learning to *reason*, but merely from learning to *pantomime rea-*

soning. LRMs technically synthesise new programs on-the-fly, but very inefficiently and shallowly.

3.5 LLM-Modulo: LLMs as an engine for creative reasoning

So, is that it? Are LLMs and LRMs a nothingburger when it comes to intelligent, creative reasoning? Well, let us not be too hasty. As we have argued, these systems fail because they lack deep understanding, lack semantics, lack grounding in the phylogeny. But what if you hooked an LLM up to something which did?

This is the key idea of Kambhampati's LLM-Modulo framework. In LLM-Modulo, an LLM (or LRM) is an engine which generates plans to solve some task, but these plans are then fed into external critics which evaluate their quality. These critiques then *backprompt* the LLM to produce better plans, until the critics are satisfied. This generate-and-test pattern echoes what philosopher Donald Campbell called "blind variation and selective retention"—the idea that all genuine knowledge acquisition, biological or otherwise, requires generating candidates without foresight and then filtering by selection Campbell 1960.

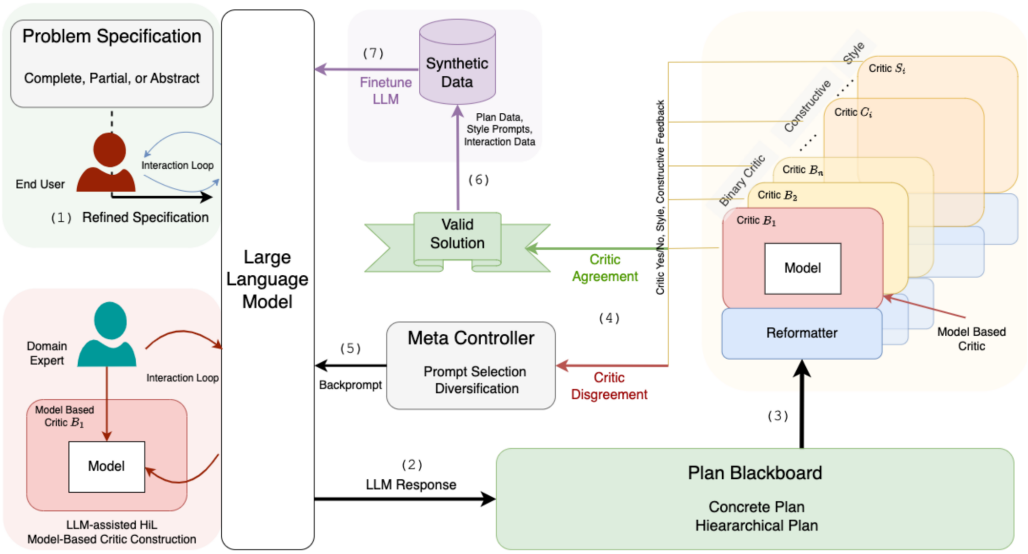


Figure 11: Kambhampati's LLM-Modulo framework: LLMs generate plans, external critics evaluate them, feedback improves outputs. Source: Kambhampati, Valmeekam, Guan, et al. 2024

These critics can ground the system. Even if to an LLM the plans are just syntax, the critics, which potentially have rich representations of the task, can thereby imbue the LLM outputs with semantics.

Does a critic make the LLM more or less creative? The answer is nuanced: more creative *within* the domain the critic enforces, but also bounded to that domain. This reflects a deeper truth we return to in our conclusions—all creativity is domain-specific, and constraints define the domain.

That was a bit abstract; let's talk about some concrete LLM-Modulo successes. One of the early breakthroughs on ARC was by AI researcher Ryan Greenblatt, who in July 2024 achieved 50% accuracy on the public version of ARC-AGI-1 by using GPT-4o to generate thousands of Python programs looking for ones which, when run, solved the training examples. This is precisely LLM-Modulo: the LLM produced plans to transform inputs

into outputs, and the critic was the Python interpreter checking whether these plans succeeded. Greenblatt also backprompted the critic's feedback into GPT-4o to revise the most promising attempts.

More recently, AI researcher Jeremy Berman (temporarily) got SOTA on ARC-AGI-2 with a similar method, but with a curious twist. Instead of producing Python code, this time the LLM (Grok-4) produced English instructions, which a sub-agent LLM then used to transform the training inputs into output grids. The critic was therefore mostly also an LLM, but there was still a final non-LLM critic to check if the 'checker' LLM's grids were correct. In December 2025, Poetiq was the first team to score over 50% on ARC-AGI-2, using Gemini 3 in a broadly similar way to Berman and Greenblatt, and beating the pure Gemini 3 Deep Think approach at half the cost. Therefore, LLM-Modulo has repeatedly demonstrated the ability to get LLMs to solve ARC puzzles significantly more accurately and efficiently.

29.4% ARC-AGI-2 (TOP SCORE!) - Jeremy Berman

Leaving behind the world of ARC, in the Summer of 2025 Google DeepMind's AlphaEvolve made a bit of a splash in the mathematics community. AlphaEvolve builds on DeepMind's earlier FunSearch (Romera-Paredes et al. 2024), which in 2024 used LLMs to discover new solutions to the cap set problem in combinatorics. AlphaEvolve is an LLM-powered coding agent which iteratively prompts an ensemble of LLMs to improve on existing programs for a given task, evaluates their performance, and then via an evolutionary algorithm uses the best-performing LLM-programs to improve the next prompt. AlphaEvolve, therefore, does LLM-Modulo.

Using AlphaEvolve, Google developed improvements to their computing ecosystem, but more excitingly discovered novel solutions to difficult mathematical problems. Their crown jewel was using AlphaEvolve to discover a novel method for multiplying 4x4 matrices in 48 multiplications (and, recursively, 16x16 in 48^2, 64x64 in 48^3, etc.) beating the record of 49 multiplications (for a recursively applicable algorithm) which had been held by Strassen's algorithm since 1969! Open-source versions like OpenEvolve and ShinkaEvolve have also been developed, with the latter discovering a SOTA method for circle-packing.

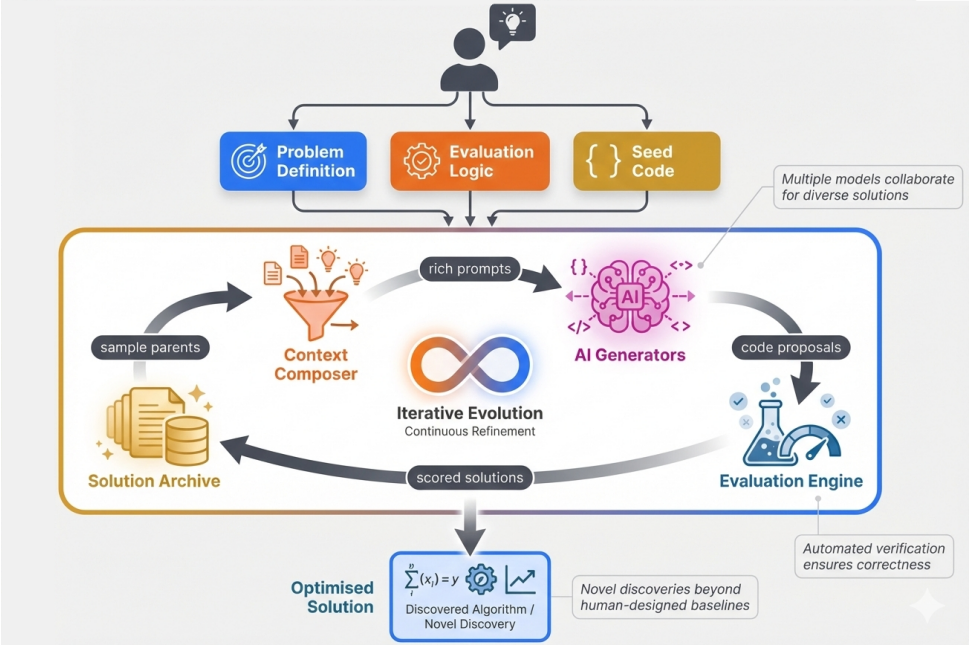


Figure 12: Summary of AlphaEvolve, generated using Nano Banana Pro based on the description from Novikov et al. 2025

Wild breakthrough on Math after 56 years... [Exclusive]

So if, as Buzzard said, LRMs have "told us nothing [mathematically] profound we didn't know already", LLM-Modulo systems like AlphaEvolve definitely have. LLM-Modulo allows these systems to be much more grounded in the phylogeny of their task, and evolutionary refinement means that these systems extend that phylogeny further. It is no coincidence that it is these systems which have produced more creative results than scaling LLMs and LRMs.

Nevertheless, these systems still rely on substantial engineering, and have so far only achieved success for narrow, well-defined tasks. To think about what that means for their creativity, let us leave LLMs behind us, and look at AlphaEvolve's older siblings...

4 Are AlphaGo and AlphaZero creative?

In March 2016, DeepMind made headlines when its AlphaGo model defeated Lee Sedol, one of the strongest players in the history of Go. Go had long been a major challenge for AI systems due to its vast depth, and until AlphaGo no AI system had ever beaten a professional player. But AlphaGo was remarkable not only in its strength, but also in the originality of some of its moves. Particularly, AlphaGo's move 37 in Game 2 amazed commentators, with Lee Sedol commenting:

> *"I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely AlphaGo is creative. This move was really creative and beautiful."*
>
> *— Lee Sedol,* AlphaGo - The Movie

AlphaGo used data from human Go games to guide its play. But its even stronger successor AlphaGo Zero used no human data at all, learning only from the rules of Go. In December 2017, DeepMind went a step further and announced AlphaZero, a more general algorithm which could learn to play many games (e.g., Go, chess, and shogi) again just from self-play, with no human data. How was this done?

4.1 Monte Carlo Tree Search

At the heart of AlphaGo, AlphaGo Zero, and AlphaZero is Monte Carlo tree search (MCTS). From every Go (or chess, or etc.) position, the possible futures of the game are a vast, ever-branching tree: each of your hundreds of possible moves makes a branch, each of your opponent's a branch of that branch, and so on, until each branch terminates in a won or lost (or drawn) position, perhaps after hundreds of moves. MCTS seeks the best path through this tree by randomly sampling many paths (in a guided way), gradually building up an accurate picture.

AlphaZero's (and AlphaGo and AlphaGo Zero's, with differences not relevant to this story) MCTS was guided by a neural network "head", which produced AlphaZero's System 1 "intuition" for how good different moves were, and how likely it was to win/lose/draw. The key training loop was to iteratively *amplify* this "intuition" via "reasoning" (i.e., MCTS), and then *distill* the conclusions of that reasoning into an enhanced "intuition". By performing MCTS guided by this neural network, and by playing against itself, AlphaZero could obtain better estimates of move quality and win probability. The neural network was then updated to better align its estimates with these improvements. Repeating this process, AlphaZero gradually climbed from random play to superhuman performance (indeed, outperforming specialised programs). AlphaZero's MCTS reasoning is vital: if one switches it off, and forces the raw "intuitive" model to play without reasoning, it plays far worse.

4.2 The creativity of AlphaGo and AlphaZero

Are AlphaGo or AlphaZero really creative, or is this an illusion? According to Rocktäschel's framework, AlphaGo is indeed open-ended:

> *"After sufficient training, AlphaGo produces policies which are novel to human expert players [...] Furthermore, humans can improve their win rate against AlphaGo by learning from AlphaGo's behavior (Shin et al., 2023). Yet, AlphaGo keeps discovering new policies that can beat even a human who has learned from previous AlphaGo artifacts. Thus, so far as a human is concerned, AlphaGo is both novel and learnable."*

The same is true of AlphaZero—in chess, AlphaZero pioneered new strategies, infamously loving to push pawns on the side of the board. AlphaGo Zero and AlphaZero are definitely not just recombining existing ideas, as they aren't given any! Unlike LLMs, who generalise somewhat by accident as a consequence of compressing their vast training data, AlphaZero plays positions it has never seen before by deliberately reasoning about them, via MCTS, and this ability was actively selected for by its training. But is this strong reasoning or weak reasoning?

There are key limits to AlphaGo/AlphaZero's reasoning. As philosopher Marta Halina highlights, the limit of AlphaGo's world is the standard game of Go; it is unable to play even mild variants of Go without retraining. Even AlphaZero, which can learn any two-player perfect-information game from its rules, can't be trained on one game and then transfer that knowledge to other games. Therefore, Halina argues that:

> *"Computer programmes like AlphaGo are not creative in the sense of having the capacity to solve novel problems through a domain-general understanding of the world. They cannot learn about the properties and affordances of objects in one domain and proceed to abstract away from the contingencies and idiosyncrasies of that domain in order to solve problems in a new context."*

Rocktäschel makes a similar point, describing AlphaGo as a "narrow superhuman intelligence" which "cannot by itself help us to discover new science or technology that requires combining insight from disparate fields".

In 2022, researchers led by Adam Gleave demonstrated an even more dramatic limit: KataGo (an even stronger Go AI than AlphaGo, developed in 2019) could be beaten a whopping 97% of the time, by using AlphaZero-style training to find *adversarial strategies* which exploited how KataGo approached the game:

> *"Critically, our adversaries do not win by playing Go well. Instead, they trick KataGo into making serious blunders that cause it to lose the game."*

The KataGo team were able to mitigate this via *adversarial training*—that is, having KataGo simulate adversarial strategies during training and learn to respond to them—but only partially. Gleave's strategies still worked 17.5% of the time even against adversarially trained KataGo; very impressive for playing Go badly!

These adversarial strategies were not arcane computer nonsense: a human expert could learn to use them to consistently beat superhuman Go AIs (and not just KataGo). Therefore, applying Rocktäschel's criteria, whilst Go AIs are "open-ended" relative to an *unassisted human observer,* relative to a human observer assisted by adversarial AI, they lack novelty in exploitable and learnable ways, and adversarial training only partially fixes this.

4.3 Does AlphaZero have phylogenetic understanding?

AlphaZero may disregard the *human* phylogeny of Go, chess, or etc., but via its self-play training loop, it creates and distills its own phylogeny: every move that it makes has a history in those millions of self-play games. This gives it some level of understanding of the moves it makes. But how sophisticated is that understanding?

A 2022 DeepMind study investigated whether AlphaZero had learned to represent human chess concepts when learning to play chess. They defined a "concept" to be a function which assigns values to chess positions (e.g., the concept of "material" adds up the value of White's pieces and subtracts the value of Black's pieces). This notion was convenient, because such functions encoding many key chess concepts have been engineered to build traditional chess programs. Using a chess database, they then trained sparse linear probes to map the activations in AlphaZero's neural network head to the functions expressing these concepts. They found that initially these probes all had very low test accuracy, but over the course of AlphaZero's training they became much more accurate for many concepts, suggesting that AlphaZero was indeed acquiring representations of those concepts. For example, after hundreds of thousands of iterations AlphaZero eventually converged on the commonly accepted values for the chess pieces.

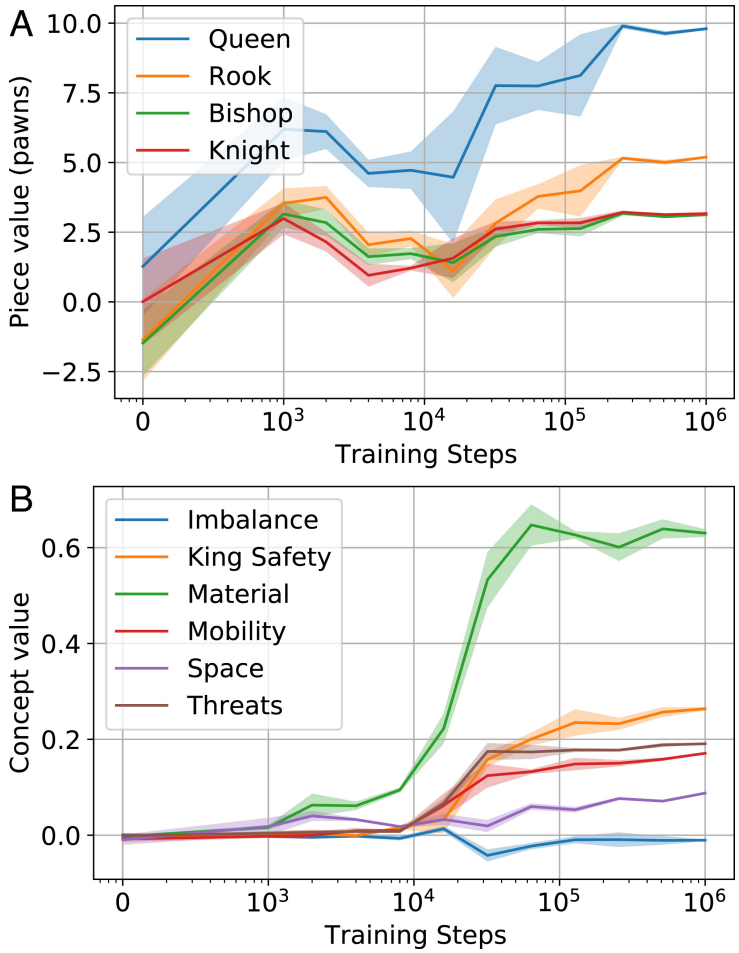


Figure 13: AlphaZero learning chess concepts over training iterations, including piece values. Source: McGrath, Kapishnikov, et al. 2022

However, there are two key caveats to this result. First, this evidence is just from sparse linear probes, which are limited tools for interpretability. Second, defining chess "concepts" as functions conflates *the positions those concepts refer to* with *what those concepts*

mean. Suppose that in all of the positions in the chess database used, in every position where someone was in check, there was never a 2x2 square all full of queens (a very rare pattern). Then both "being in check" and "being in check with no 2x2 square of queens on the board" would correspond to the same function, but obviously don't mean the same thing.

As highlighted by Fodor and Pylyshyn (in their classic critique of connectionism), understanding these *meanings* requires grasping their *systematic* and *compositional* nature. Understanding "being in check" should be intrinsically tied up with understanding "being in check by a pawn", "blocking a check", "pinning a piece to the King" etc. The research does not explore such networks of interrelated understandings, and as such cannot demonstrate a deep abstract understanding of these concepts.

But is abstract understanding needed to understand chess (or Go)? Abstract explanations might at first seem like the pinnacle of understanding, but the development of chess theory (as described for example in John Watson's classic *Secrets of Modern Chess Strategy*) has increasingly favoured concrete explanations of chess positions. For example, in his 2025 article "Understanding and Knowing", Grandmaster Matthew Sadler describes three levels of understanding of an example chess position, which get more and more concrete.

From the concrete point of view, to understand a chess position is *not* to identify abstract principles or concepts which explain the winning strategy, but rather to grasp a sample of the game tree which highlights all of the key variations (i.e., the winning line, the refutations of the various alternatives, etc.) and contrasts those with key variations in similar positions. But this sample must not be *exhaustive*: a brute-force description of the entire tree (besides taking the lifetime of the universe to compute) would not showcase understanding. Concrete understanding means only needing to look at the key lines, and in recognising how sensitive (or not) those lines are to the specific features of the position.

On these concrete terms, AlphaZero represents some progress, as it looks at far fewer positions than older chess systems in its search. However, it still looks at thousands of times more positions than a human grandmaster, so it will still explore many irrelevant lines. More deeply, it lacks *counterfactual* understanding: in the above Sadler article, a crucial piece of the highest level of understanding was seeing how (and why) the winning line in the position *wouldn't work* in a superficially similar position. AlphaZero's MCTS will never explore these sorts of counterfactual positions. The adversarial examples show that even in purely concrete terms, these systems can utterly fail to understand strange positions.

In summary, AlphaGo and AlphaZero are more creative than LLMs (and LRMs), due to having a degree of concrete understanding. Move 37 was an authentic creative discovery, originating in AlphaGo. They are capable of broad exploration within the domain of the rules they are given. In Boden's terms, they exhibit exploratory creativity—discovering new possibilities within their conceptual space—but lack the transformational creativity needed to extend beyond it. Ultimately, their understanding still suffers from limitations which hamper their creativity, and make them less robust and unable to generalise beyond the game they learned. Despite their immense strength, they are blind to deeper domain-general features, and can be bamboozled by spurious patterns even within the domain they were trained on.

But if AlphaZero lacks domain-general understanding yet crushes humans, does understanding matter? The answer depends on what you want. AlphaZero is stronger than any human at chess—but would fail at "chess with one rule change" without retraining from scratch. Carlsen, though weaker, could adapt instantly, and would be very hard to beat by playing badly! For robust generalisation to unknown unknowns, the deeper understand-

ing matters; for raw performance on a fixed task, it may not. DeepMind's work does suggest that AlphaZero learned to represent key chess concepts—AlphaZero shows non-zero understanding—but did not demonstrate a deep, systematic, compositional understanding of these concepts.

5 Putting the humans back in the loop

We have now surveyed the landscape of autonomous AI creativity: LLMs and LRMs interpolate their training data without deep understanding; AlphaGo and AlphaZero reason more authentically within narrow domains but cannot generalise beyond their given rules. None of these systems, operating alone, can match the "unknown unknown" creativity that characterises human intelligence.

But have we been asking the wrong question this whole time? So far, we have been focusing on whether AI systems, by themselves, can reason creatively. This framing echoes the dream (or nightmare, depending on who you ask) of fully autonomous AI systems, a dream infamously expressed by Nobel laureate Geoffrey Hinton in 2016:

> *"I think if you work as a radiologist you're like the coyote that's already over the edge of the cliff but hasn't yet looked down so doesn't realize there's no ground underneath him. People should stop training radiologists now. It's just completely obvious that within 5 years deep learning is going to do better than radiologists because it's going to be able to get a lot more experience. It might be 10 years but we've got plenty of radiologists already."*
>
> *Geoff Hinton: On Radiology*

History has not been kind to this prediction, but setting aside the inaccuracy of the timeline, notice how Hinton pictures deep learning as *replacing* radiologists, rendering them obsolete. But what if instead the future looks like radiologists and AI systems *working together*, to perform better than either could alone, or do radiology in more diverse settings? Then there might be a need for more radiologists than ever. Spreadsheets, after all, did not lead to fewer accountants.

The bedrock of contemporary medical imaging is CT and MRI scanning; fantastic techniques, but which use large, expensive machines, rendering them inaccessible to people in remote or poorer locations. In sub-Saharan Africa, there is less than one MRI scanner per million people, with many countries having none at all; in the USA, there are nearly 40 MRI scanners per million people. In recent years, AI methods have significantly enhanced new medical imaging techniques, e.g. photoacoustic imaging, which have the potential to be cheaper and more portable alternatives to MRI—but which still need radiologists! There is hope that these new techniques could significantly expand the reach of medical imaging in places such as Africa. If that hope is realised, this could dramatically increase global demand for radiologists.

In terms of creative reasoning, we should therefore be thinking not only about AI creativity, but also human-AI co-creativity. Consider coding and science; these are inherently *interactive* endeavours: any AI system will inevitably be interfacing with humans throughout these tasks. Who commissioned the software? Who are its users? Who will perform AI-designed experiments? To quote the AlphaEvolve authors from our MLST interview:

> *"I think the thing that makes AlphaEvolve so cool and powerful is kind of this back and forth between humans and machines, right? And like, the humans ask questions. The system gives you some form of an answer. And then you, like, improve your intuition. You improve your question-asking ability, right? And you ask more questions. [...] We're exploring [the next level of human-AI interaction] a lot. And I think it's very exciting to see, like, what can be done in this kind of symbiosis space."*
>
> *Wild breakthrough on Math after 56 years... [Exclusive]*

DeepMind researchers Mathewson and Pilarski highlight how humans are embedded throughout the machine learning lifecycle, from data collection to deployment. The *Neuroevolution* textbook echoes this too: "humans and machines can work synergistically to construct intelligent agents," ultimately enabling "interactive neuroevolution where human knowledge and machine exploration work synergistically in both directions to solve problems". We have so far been focusing on the "I" of AI, but the "A" often hides the extensive reliance of these systems on humans.

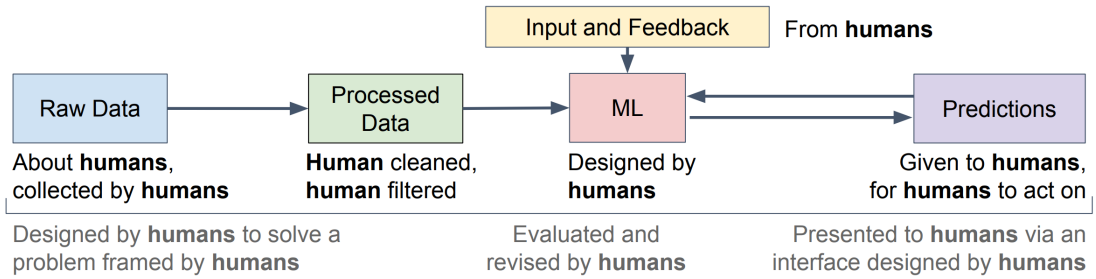


Figure 14: All machine learning is interactive: humans are embedded throughout the AI development lifecycle. Source: Mathewson and Pilarski 2022

Consider some famous examples. The deep learning revolution, so the story goes, was kicked off by the success of the convolutional neural network AlexNet in the 2012 ImageNet challenge. Datasets like ImageNet have been instrumental in the success of deep learning, yet ImageNet would not exist without massive human labour—an untold number of Amazon Mechanical Turk users worked for over two years to label the 14 million images. ChatGPT spawned the LLM era, and unlike AlexNet it was trained via *self-supervised learning*: using the training data itself to label the next token, without needing explicit human labelling. Yet this training data—essentially, a large chunk of the internet—was of course created by humans. Making ChatGPT presentable also required extensive *reinforcement learning with human feedback*. A TIME investigation found that this relied on significant (and traumatising) Kenyan labour.

Will AI always rely on human labour? Could not future AI systems be trained on AI-generated data and supervised by AIs, without any humans in the loop? Anthropic have after all been pioneering reinforcement learning with *AI* feedback, and the big tech companies have reportedly turned to synthetic data because they are running out of internet to train on. However, a 2024 front-page *Nature* paper warned that indiscriminately training AIs on AI-generated data leads to "model collapse"—an irreversible disappearance of the tails (i.e., low-probability outputs) of the AI's distribution. This would be especially fatal for creativity, since losing the tail means losing unexpected and novel outputs. Human-AI collaborations can exploit complementary strengths: humans often find generation harder than evaluation, whilst AI systems often demonstrate the reverse. Thus, by del-

egating tasks, such as in LLM-Modulo, one can get the best of both worlds. As Stanley argues, the human ability to recognise interestingness is irreplaceable:

> *"We have a nose for the interesting. That's how we got this far. That's how civilization came out. That's why the history of innovation is so amazing for the last few thousand years."*
>
> *Prof. KENNETH STANLEY - Why Greatness Cannot Be Planned*

5.1 What does human-AI co-creativity look like?

In 1997, Deep Blue beat chess world champion Garry Kasparov, and by 2006 computers had decisively overtaken human chess players—Hydra crushed Michael Adams 5½–½ in 2005, and Deep Fritz beat world champion Vladimir Kramnik 4–2 in 2006. (AlphaZero would later join the party with a bang in 2017.) As we saw, Go went the same way in 2016. Human-AI collaboration is now an integral part of high-level play in both games, with top players extensively preparing with computers. One might worry that this would atrophy these players' creative minds, but quite the opposite seems true. After the advent of AlphaGo, human Go players began to play both more accurately and more creatively. This really kicked in when *open-source* superhuman Go AIs arrived, as people could then learn not only from the AIs' *actions*, but also from their *reasoning processes*.

A similar story is true of chess: not only do players play much more accurately now than in the past, but computer analysis helped overturn dogmatic ideas of how chess could be played, and breathed new life into long abandoned strategies. AlphaZero has been used to explore new variant rules for chess, dramatically faster than humans could alone. Most recently, in a 2025 paper DeepMind showed how chess patterns uniquely recognised by AlphaZero could be extracted and taught to human grandmasters, demonstrating that these systems can continue to enhance the human understanding of chess.

AlphaZero in Chess | Reflections on Creative Play

Moving beyond the world of board games, let us once again consider Stanley's Picbreeder project. Picbreeder was a deeply unorthodox way of training a generative AI. No training data and SGD and all that; instead Picbreeder embedded a generative architecture within a community of users, and used their social dynamics and aesthetic tastes to evolve the parameters of that architecture. This allowed for *open-ended exploration* of the space of

images this architecture could produce, far beyond what standard training algorithms can achieve.

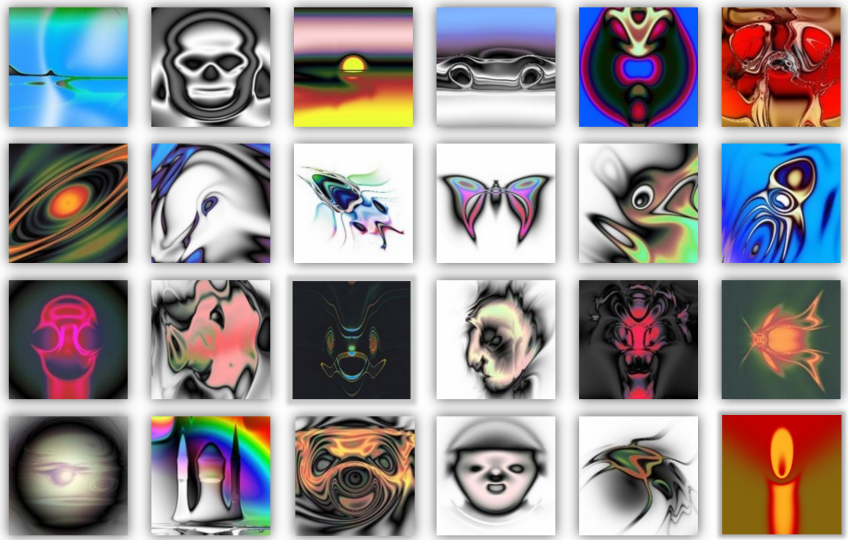


Figure 15: Picbreeder networks learn semantically meaningful representations through open-ended evolution. Source: Stanley 2014

As we keep mentioning, this process produced vastly superior representations to the ones they got by naively training (by SGD) the same architecture to produce that target image. Now, their experiment wasn't exactly fair: that training method is a bit *too* naive, since to learn to make one image there is no need to learn any general representations—just memorise the image. Complex representations would be wasted effort. To make a fairer comparison, diffusion models (when trained on more images than they can memorise) do learn interesting representations. Nevertheless, the representations learned by Picbreeder networks possess semantic information on a whole other level. For the "skull", they described eye and mouth shape, winking, and opening the mouth. For the "butterfly", they described colour, wing size, and even converted it to a fly. These remarkable properties show what you can get if you keep humans in the loop.

Finally, human-AI collaborations may also soon be fruitful in academia. Or so argued Fields medallist Terence Tao in a 2024 interview for *Scientific American*. Inspired by the success of automated proof assistants like Lean, Tao imagines mathematicians and AIs soon working together to produce proofs:

> *"I think in three years AI will become useful for mathematicians. It will be a great co-pilot. You're trying to prove a theorem, and there's one step that you think is true, but you can't quite see how it's true. And you can say, 'AI, can you do this stuff for me?' And it may say, 'I think I can prove this.'"*

Tao suggests that this might, eventually, radically change the way that mathematics is done, from mathematicians as "individual craftsmen" to a mass-production pipeline, "proving hundreds of theorems or thousands of theorems at a time", with human mathematicians directing the AIs at a higher level. Tao is optimistic about the effects of AI on mathematics. Asked "So instead of this being the end of mathematics, would it be a bright future for mathematics?" Tao replied that AI would diversify mathematical practice, creating "different ways of doing mathematics that just don't exist right now". Tao closed

with a particularly interesting insight on the power of human-AI collaboration:

> *"So much knowledge is somehow trapped in the head of individual mathematicians. And only a tiny fraction is made explicit. But the more we formalize, the more of our implicit knowledge becomes explicit. So there'll be unexpected benefits from that."*

6 The Structure of Creativity

6.1 The Semantic Graph

Having spent two years surveying the landscape of AI creativity, these are some of our distilled thoughts and arguments. There is an interesting distinction between what we might call *statistical creativity* and *semantic creativity*. Statistical creativity—making it more likely that we stumble upon interesting regions of a search space—is what current AI systems excel at. But genuine creativity may require something more: a pure form where we are in posession of the semantic graph, where our steps are constrained by deep structural knowledge rather than probabilistic guesswork. As Risi, Tang, Ha, and Miikkulainen argue, "neuroevolution gives us a rare opportunity to study representations not just as a byproduct of loss minimization, but as artifacts of open-ended exploration and accumulated structural regularities" Risi et al. 2025. As one of us (TS) put it in conversation with Risto Miikkulainen:

> *"We are describing a kind of statistical creativity where we want to make it more likely that we will find these tenuous, interesting regions. But could there be a kind of almost pure form of creativity where we know the semantic graph?"*

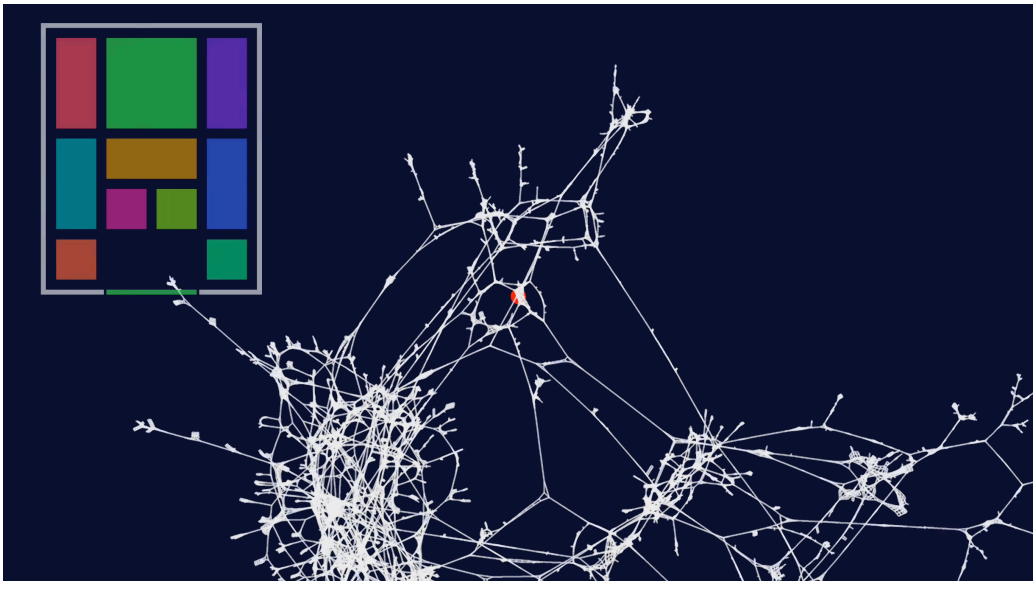


Figure 16: The state space of the Klotski sliding block puzzle, visualised as a graph. Each node is a board configuration; edges connect positions one move apart. The graph reveals distinct substructures—local regions with their own logic—connected by tenuous paths. Creativity, in this view, is finding those tenuous connections that lead to entirely different substructures. Source: 2swap 2025

We think this visualisation is a powerful intuition pump: it shows what a fully known "semantic graph" would look like. In real creative domains, of course, the space is only partially observable—we discover new dimensions and subspaces as we go, rather than navigating a known topology. This is what Stanley means when he describes transformational creativity as "adding new dimensions to the universe"—not just finding a new location within a known space, but discovering entirely new subspaces. The stepping stones that matter are those tenuous connections between clusters that open up previously unknown regions. As Miikkulainen put it when shown this visualisation: creativity involves "pushing into another area of kind of solutions that you've never seen before" by finding those rare transitions between substructures.

But how do we measure creativity? Perhaps the answer lies in the *size* of the subspace discovered. The stepping stone that leads to a vast new region of possibilities is more creative than one that leads to a small cul-de-sac. Looking at the Klotski graph, we can immediately see which clusters are large and which connections are most valuable—but this is a fully observable domain. In the real world, we operate under a "fog of war": we discover the stepping stones as we go, and it takes time to realise just how big a new subspace actually is. Only in retrospect, once the phylogenetic tree has been expanded by subsequent discoveries, can we recognise that an intermediate stepping stone was extraordinarily creative.

Can we tell *in advance* whether a stepping stone will be creative? Akarsh Kumar suggests the answer lies in *evolvability*—the capacity to enable future discoveries:

> *"There's an implicit selection pressure for evolvable things. If there's two versions of the skull—one is spaghetti and one is modular and composable—after a few generations of evolution, the one that's more evolvable will win out. Just like in natural evolution, the evolution of evolvability. And this evolvability combined with serendipity is what gives you these nice representations."*
>
> *AI is SO Smart, Why Are Its Internals 'Spaghetti'?*

This is why path-dependent representations matter: they encode not just solutions, but *potential*—the latent capacity for future creative leaps.

This echoes what Akarsh Kumar calls the difference between "statistical intelligence" and "regularity-based intelligence"—the former perfect at pattern matching, the latter grounded in the actual structure of the world (MLST interview). Statistics are wonderful for building representations of data, for memorising what already exists. But intelligence—and creativity—may be fundamentally about building *new* representations, new models, constrained by the path that got us there.

6.2 Constraints as Enablers

But constraints are not the enemy of creativity—they are its very foundation. As Noam Chomsky argued in our interview:

> *"In fact, while it's true that our genetic program rigidly constrains us, I think the more important point is that the existence of that rigid constraint is what provides the basis for our freedom and creativity. [...] If we really were plastic organisms without an exten-*

> *sive preprogramming, then the state that our mind achieves would in fact be a reflection of the environment, which means it would be extraordinarily impoverished. Fortunately for us, we're rigidly preprogrammed with extremely rich systems that are part of our biological endowment. Correspondingly, a small amount of rather degenerate experience allows a kind of a great leap into a rich cognitive system. [...] We can say anything that we want over an infinite range. Other people will understand us, though they've heard nothing like that before. We're able to do that precisely because of that rigid programming."*

The Ghost in the Machine (01:25:33)

Neuroevolution pioneer Risto Miikkulainen affirmed this insight in conversation with one of us (TS): "Isn't it weird that to be more creative, you actually have to respect the constraints in the phylogeny more? Because you think of creativity as being like, I can discover anything. But it's actually kind of the opposite." Miikkulainen agreed: "Yeah. It's respecting the constraints of the problem."

All creativity respects constraints—but deep understanding of a domain's constraints often enables transfer to related domains, precisely because it captures structural regularities rather than surface features. This is why Carlsen can play chess variants while AlphaZero cannot: Carlsen grasps the deep structure of chess, not merely its surface presentation. An unconstrained system would be less coherent, not more creative. Creativity without constraints is not creativity at all—it is noise.

The late Margaret Boden crystallised this insight in her landmark study of creativity. "Far from being the antithesis of creativity," she wrote, "constraints on thinking are what make it possible" Boden 2004. This is because "constraints map out a territory of structural possibilities which can then be explored, and perhaps transformed to give another one". The alternative is not freedom but chaos: "to drop all current constraints and refrain from providing new ones is to invite not creativity, but confusion". Boden doesn't mince words: "there, madness lies". Examining the great creative minds of history, Boden observed that "they respect constraints more than we do, not less"—they are more free precisely because they understand the domain's constraints well enough to push beyond them.

Here is one way to think about it: creativity is like building a jigsaw puzzle you never knew existed. Ordinary puzzle-solving means fitting pre-cut pieces into a picture you already have on the box. But creative work means discovering the pieces *as* you discover the picture—each new piece reshapes what the whole might become. You cannot interpolate your way to a picture you have never seen. You must feel your way forward, constrained

by the pieces you have found so far, until the image emerges.

6.3 The Supervisor Illusion

Epistemologists have long distinguished between *knowing* isolated facts and *understanding* how they cohere. Understanding requires "grasping of explanatory and other coherence-making relationships", not just believing isolated pieces of information Baumberger, Beisbart, and Brun 2017. A child who learns "greenhouse gases cause warming" via testimony knows the explanation but does not understand it—he cannot answer counterfactual questions or reason about the mechanism. Current AI systems are in the same position: they can reproduce explanations without possessing the coherence that would make those explanations *understood*.

"AI slop" represents exactly this—to paraphrase Boden: there, slop lies, the opposite of coherence and therefore the opposite of creativity. Slop is what happens when an artifact is generated without path dependence, without understanding, without respecting the phylogeny or the constraints. As Kumar and Stanley argued in their work on fractured and entangled representations Kumar, Clune, et al. 2025, LLM outputs are incoherent precisely because they took the wrong path—or rather, no coherent path at all. Their representations lack the stepping-stone structure that would make outputs meaningful. It is only possible to make language models produce non-slop when they are guided by a competent supervising human who provides the missing coherence.

There is a curious asymmetry here worth noting. Language models in *generation* mode are far more likely to produce slop than when operating in *discrimination* mode. Consider that the same LLM that confidently hallucinates a citation when asked to generate one can, when prompted to verify that citation, correctly identify it as nonexistent. Why can AI detect its own hallucinations but not avoid generating them? We suspect the answer lies in the nature of the two tasks. In discrimination mode, the model is given a specific, constrained task: does this text exhibit certain statistical signatures? The constraints of the task impose coherence. In generation mode, the model must conjure coherence from nothing, and without external guidance it defaults to the statistically average—which is to say, the mediocre, the derivative, the slop. This explains why agentic workflows that decompose generation into smaller, more constrained subtasks—like verifying each reference individually rather than generating a bibliography in one shot—can dramatically reduce hallucination and improve coherence. The constraints of the subtask substitute for the understanding the model lacks.

This pattern extends beyond citation checking. With increasing levels of specification, and in domains where outputs are verifiable—even implicitly verifiable through execution or compilation—language models perform dramatically better. Tools like Claude Code, and indeed most of the recent practical advances in deploying large language models, are fundamentally ways of adding constraints to the generation process. Agentic scaffolding, tool use, code execution, test suites, type systems: all of these impose external structure that guides generation toward coherence. In effect, we are compensating for the models' lack of phylogenetic understanding by adding constraints that make them *act as if* they had such understanding. The constraints do the work that deep structural knowledge would otherwise provide.

This creates what we might call the *supervisor illusion*. When a competent expert uses an AI system, they implicitly provide the constraints that guide generation toward coherence—through precise prompts, iterative refinement, and knowing which outputs to reject. The result can be impressive, and it is tempting to credit the AI with creativity it does not pos-

sess. This illusion is particularly seductive in Silicon Valley, where technically sophisticated users routinely coax remarkable outputs from AI systems and extrapolate to world-changing predictions. As Melanie Mitchell has argued, most AI benchmarks lack *construct validity*—they fail to predict real-world performance because impressive results often stem from data contamination, approximate retrieval, or exploitable shortcuts rather than genuine capability Mitchell 2026. Anthropic CEO Dario Amodei, for instance, recently suggested that AI could "displace half of all entry-level white collar jobs in the next 1–5 years" while enabling "10–20% sustained annual GDP growth" Amodei 2025. In the same essay, Amodei notes that "top engineers now delegate almost all their coding to AI"—but this inadvertently proves the point: it is precisely because they are *top engineers* that the delegation works. They provide the missing coherence, acting as the verifier in an LLM-Modulo loop, rendering the AI's statistical outputs into genuine solutions. There is a second factor here too: top engineers can move fast with AI-generated code because they comprehend what is happening—they can breeze through the process without incurring what we might call *understanding debt*. When less experienced engineers attempt the same velocity, they outpace their own comprehension. The code works (for now), but they do not understand why, and this debt compounds. Every shortcut becomes a liability when something breaks. This is why extrapolations from expert productivity to market-wide transformation are likely to disappoint.

But such projections also assume AI performs consistently across users. When people with limited domain expertise use the same systems, slop is the inevitable result—not because the AI has become less capable, but because the supervising intelligence that was secretly doing the heavy lifting is no longer present.

To be fair, modern AI does raise the floor: even naive users can produce *some* level of coherence that would have been difficult before. But the ceiling remains firmly dependent on the supervisor. The gap between floor and ceiling is precisely the gap between statistical pattern-matching and genuine understanding. AI amplifies what you bring to it; it does not substitute for what you lack.

6.4 Intelligence Without Understanding

As we noted at the outset, intelligent reasoning needs creativity—but creativity doesn't need intelligence. Evolution produced the entire tree of life through blind variation and selective retention, with no understanding at all. But can raw intelligence substitute for understanding? We are sceptical. A superintelligence that lacked domain knowledge would be no better than the million monkeys, just faster at trying random paths. As Philip K. Dick put it, "reality is that which, when you stop believing in it, doesn't go away" Dick 1978—and without access to reality's constraints, no amount of raw cognitive power will help you explore it.

This may explain why creative solutions often seem obvious in hindsight—what we might call the "McCorduck effect" for creativity. The path that was tenuous becomes a well-worn road. But perhaps the obviousness is real: genuine creativity follows the constraints of the domain, and the solution *was* always latent in the structure, waiting to be discovered by someone who understood it deeply enough to find the tenuous connection. This applies to what Boden calls *exploratory* creativity—navigating within an existing conceptual space. Exploratory ideas, she notes, "may come to seem glaringly obvious ('Ah, what a foolish bird I have been!')" Boden 2004. But *transformational* creativity is different: it breaks the space itself, generating what Boden calls "impossibilist surprise":

> *"Where transformational creativity is concerned, the shock of the new may be so great that even fellow artists find it difficult to see value in the novel idea."*

Quantum mechanics still feels strange, not because we haven't understood it, but because classical intuitions cannot be patched to include it. The prior conceptual space has been broken, not extended.

7 Conclusions

Our central claim is this: current AI systems lack the structured understanding required for genuine creativity. LLMs interpolate their training data; they can recombine existing patterns but cannot navigate to regions that require understanding constraints they have never explicitly encoded. AlphaZero reasons within narrow domains but cannot transfer that understanding to novel contexts. Neither possesses the path-dependent representations that encode not just solutions but *evolvability*—the capacity to enable future discoveries. Without such representations, without access to the coherence-making relationships that constitute understanding, AI systems produce statistical creativity at best: making it more likely we stumble upon interesting regions, but never navigating there deliberately.

This is not a claim that AI can *never* be creative. The philosophical debate remains open. If functionalism is correct—if what matters is the computational structure rather than the substrate—then AI systems with genuine structured understanding might one day be possible. The question of grounding looms large: can a system that has never pushed against reality's constraints ever truly understand them? Perhaps future AI systems, trained through interaction with the physical world rather than passive consumption of text, could develop the kind of understanding that current systems lack. Perhaps they could achieve creativity in domains we cannot access, even if not in ours. We do not rule this out.

But for now, the most promising path forward is human-AI co-creativity. Chess and Go players have become more creative, not less, by working with superhuman AI systems. Picbreeder showed how keeping humans in the loop can produce representations far richer than those achieved by standard training methods. And as Terence Tao suggests, mathematicians and AI systems working together may soon prove theorems that neither could reach alone. The human provides the coherence, the understanding, the nose for the interesting; the AI provides statistical power, tireless exploration, and freedom from cognitive biases. Together, they can explore vast spaces of possibility that neither could navigate alone.

If greatness cannot be interpolated, perhaps it cannot be fully automated either—at least not yet. But it can be amplified.

We are producing some upcoming videos on these topics. Stay tuned.

References

2swap (2025). *I Solved Klotski*. YouTube. Visualization of the Klotski puzzle's state space as a graph, showing how local substructures connect via tenuous paths. URL: https://www.youtube.com/watch?v=YGLNyHd2w10.

Amodei, Dario (2025). *The Adolescence of Technology*. Personal Essay. Anthropic CEO's predictions on AI economic impact and job displacement. URL: https://www.darioamodei.com/essay/the-adolescence-of-technology.

Baumberger, Christoph, Claus Beisbart, and Georg Brun (2017). "What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science". In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Ed. by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon. Distinguishes knowledge (acquirable through testimony) from understanding (requiring grasp of coherence-making relationships). New York: Routledge, pp. 1–34. ISBN: 978-1138921931.

Beger, Claas, Ryan Yi, Shuhao Fu, Arseny Moskvichev, Sarah W. Tsai, Sivasankaran Rajamanickam, and Melanie Mitchell (2025). "Do AI Models Perform Human-like Abstract Reasoning Across Modalities?" In: *arXiv preprint arXiv:2510.02125*. Tests LRMs on ConceptARC benchmark. arXiv: 2510.02125 [cs.AI]. URL: https://arxiv.org/abs/2510.02125.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. New York, NY, USA: Association for Computing Machinery, pp. 610–623. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

Bird, Jon and Paul Layzell (2002). "The Evolved Radio and its Implications for Modelling the Evolution of Novel Sensors". In: *Proceedings of the 2002 Congress on Evolutionary Computation (CEC 2002)*. IEEE, pp. 1836–1841. DOI: 10.1109/CEC.2002.1004522. URL: https://people.duke.edu/~ng46/topics/evolved-radio.pdf.

Boden, Margaret A. (2004). *The Creative Mind: Myths and Mechanisms*. 2nd. London: Routledge. ISBN: 978-0415314534.

— (2009). "Computer Models of Creativity". In: *AI Magazine* 30.3. Overview of computational creativity including response to Lovelace's objection, pp. 23–34. DOI: 10.1609/aimag.v30i3.2254. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2254.

Bonnaire, Tony, Aurélien Decelle, Davide Ghio, Giulio Biroli, Cédric Fevotte, and Lenka Zdeborová (2025). "Why Diffusion Models Don't Memorize: The Role of Implicit Dynamical Regularization in Training". In: *arXiv preprint arXiv:2505.17638*. Shows that global minimisers of diffusion model objectives would perfectly memorise, but training dynamics prevent this. NeurIPS 2025 Best Paper Award. arXiv: 2505.17638 [cs.LG]. URL: https://arxiv.org/abs/2505.17638.

Bubeck, Sébastien (2025). *Claim: gpt-5-pro can prove new interesting mathematics*. Twitter/X. Tweet claiming GPT-5-pro proved a better bound than a convex optimization paper. URL: https://x.com/SebastienBubeck/status/1958198661139009862.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang (2023). "Sparks of Artificial General Intelligence: Early experiments with GPT-4". In: *arXiv preprint arXiv:2303.12712*. arXiv: 2303.12712 [cs.CL]. URL: https://arxiv.org/abs/2303.12712.

Burnham, Greg (2025). *We Didn't Learn Much from the IMO*. Epoch AI Gradient Updates. Analysis of LRM performance on the 2025 International Mathematical Olympiad. URL: https://epoch.ai/gradient-updates/we-didnt-learn-much-from-the-imo.

Campbell, Donald T. (1960). "Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes". In: *Psychological Review* 67, pp. 380–400. DOI: 10.1037/h0040373.

Chollet, François (2019). "On the Measure of Intelligence". In: *arXiv preprint arXiv:1911.01547*. arXiv: 1911.01547 [cs.AI]. URL: https://arxiv.org/abs/1911.01547.

— (2024). *Four Levels of Generalization*. Twitter/X. Tweet describing four levels of generalization in AI systems. URL: https://x.com/fchollet/status/1763692655408779455.

Chomsky, Noam (2023). *The Ghost in the Machine – Noam Chomsky*. Machine Learning Street Talk (YouTube). Interview exploring language, cognition, and how constraints provide the basis for creativity. URL: https://www.youtube.com/watch?v=axuGfh4UR9Q.

Dick, Philip K. (1978). *How to Build a Universe That Doesn't Fall Apart Two Days Later*. Lecture/Essay. Contains the famous quote: "Reality is that which, when you stop believing in it, doesn't go away". URL: https://urbigenous.net/library/how_to_build.html.

Fodor, Jerry A. and Zenon W. Pylyshyn (1988). "Connectionism and Cognitive Architecture: A Critical Analysis". In: *Cognition* 28.1-2, pp. 3–71. DOI: 10.1016/0010-0277(88)90031-5. URL: https://www.sciencedirect.com/science/article/pii/0010027788900315.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (2014). "Explaining and Harnessing Adversarial Examples". In: *arXiv preprint arXiv:1412.6572*. arXiv: 1412.6572 [stat.ML]. URL: https://arxiv.org/abs/1412.6572.

Google DeepMind (2025). "AlphaEvolve: A coding agent for scientific and algorithmic discovery". In: *Google DeepMind Blog*. Technical blog post describing AlphaEvolve system. URL: https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/.

Halina, Marta (2021). "Insightful Artificial Intelligence". In: *Mind & Language* 36.3, pp. 315–329. DOI: 10.1111/mila.12321. URL: https://onlinelibrary.wiley.com/doi/10.1111/mila.12321.

Hughes, Edward, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktäschel (2024). "Open-Endedness is Essential for Artificial Superhuman Intelligence". In: *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. Vol. 235. Proceedings of Machine Learning Research. Formal definition of open-endedness based on novelty and learnability. PMLR. arXiv: 2406.04268 [cs.AI]. URL: https://arxiv.org/abs/2406.04268.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux. ISBN: 978-0374275631.

Kambhampati, Subbarao (2024). *LLMs Don't Reason, They Memorize: Subbarao Kambhampati (ICML 2024)*. Machine Learning Street Talk (YouTube). Interview on LLM planning limitations and the LLM-Modulo framework. URL: https://www.youtube.com/watch?v=y1WnHpedi2A.

Kambhampati, Subbarao, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy (2024). "LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks". In: *arXiv preprint arXiv:2402.01817*. arXiv: 2402.01817 [cs.AI]. URL: https://arxiv.org/abs/2402.01817.

Kambhampati, Subbarao, Karthik Valmeekam, Atharva Gundawar, Daman Arora, Lin Guan, Kaya Stechly, and Mudit Verma (2025). "Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces!" In: *arXiv preprint arXiv:2504.09762*. Critique of anthropomorphising LLM reasoning tokens. arXiv: 2504.09762 [cs.AI]. URL: https://arxiv.org/abs/2504.09762.

Kohs, Greg (2017). *AlphaGo*. Documentary Film. Documentary following DeepMind's AlphaGo and the match against Lee Sedol. URL: https://www.alphagomovie.com/.

Kuhn, Thomas S. (2012). *The Structure of Scientific Revolutions*. 50th Anniversary. Originally published 1962. Chicago: University of Chicago Press. ISBN: 978-0226458120.

Kumar, Akarsh, Jeff Clune, Joel Lehman, and Kenneth O. Stanley (2025). "Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis". In: *arXiv preprint arXiv:2505.11581*. Shows Picbreeder networks have remarkably well-structured representations compared to SGD-trained networks. arXiv: 2505.11581 [cs.NE]. URL: https://arxiv.org/abs/2505.11581.

Kumar, Akarsh and Tim Scarfe (2024). *AI, Evolution, and Path-Dependent Representations*. Machine Learning Street Talk (MLST) Podcast. Discussion of statistical vs regularity-based intelligence and the FER hypothesis. URL: https://youtu.be/MutwZIJKnj4.

Legg, Shane and Marcus Hutter (2007). "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds and Machines* 17.4, pp. 391–444. DOI: 10.1007/s11023-007-9079-x. URL: https://arxiv.org/abs/0712.3329.

Lehman, Joel and Kenneth O. Stanley (2011). "Abandoning Objectives: Evolution Through the Search for Novelty Alone". In: *Evolutionary Computation* 19.2, pp. 189–223. DOI: 10.1162/EVCO_a_00025. URL: https://www.cs.swarthmore.edu/~meeden/DevelopmentalRobotics/lehman_ecj11.pdf.

Lorenz, Edward N. (1963). "Deterministic Nonperiodic Flow". In: *Journal of the Atmospheric Sciences* 20.2, pp. 130–141. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml.

Lu, Ximing, Melanie Sclar, Skyler Hallinan, Faeze Brahman, Liwei Jiang, Jaehun Jung, Peter West, Alane Suhr, Ronan Le Bras, and Yejin Choi (2025). "AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text". In: *The Thirteenth International Conference on Learning Representations (ICLR 2025)*. Introduces the Creativity Index metric for measuring linguistic creativity. CC-BY license. URL: https://openreview.net/forum?id=ilOEOIqolQ.

Mathewson, Kory W. and Patrick M. Pilarski (2022). "A Brief Guide to Designing and Evaluating Human-Centered Interactive Machine Learning". In: *arXiv preprint arXiv:2204.09622*. Argues that humans are embedded throughout the AI development lifecycle. arXiv: 2204.09622 [cs.LG]. URL: https://arxiv.org/abs/2204.09622.

McCorduck, Pamela (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. 2nd. Originally published 1979. Documents the "AI effect": the tendency to dismiss AI achievements as "not really intelligence" once accomplished. Natick, MA: A.K. Peters. ISBN: 978-1568812052.

McGrath, Thomas, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik (2022). "Acquisition of chess knowledge in AlphaZero". In: *Proceedings of the National Academy of Sciences* 119.47, e2206625119. DOI: 10.1073/pnas.2206625119. URL: https://doi.org/10.1073/pnas.2206625119.

McGrath, Thomas, Nenad Tomašev, Matthew Sadler, Natasha Regan, David Silver, and Demis Hassabis (2025). "Bridging the human-AI knowledge gap through concept discovery and transfer in AlphaZero". In: *Proceedings of the National Academy of Sciences*. Demonstrates extracting AlphaZero chess patterns to teach human grandmasters. DOI: 10.1073/pnas.2406675122. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2406675122.

Mitchell, Melanie (2019). *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux. ISBN: 978-0374257835.

— (2026). *On Evaluating Cognitive Capabilities in Machines (and Other "Alien" Intelligences)*. AI Guide (Substack). Discusses construct validity: AI benchmarks fail to predict real-world performance. URL: https://aiguide.substack.com/p/on-evaluating-cognitive-capabilities.

MLST (2025). *Google AlphaEvolve – Discovering New Science (Exclusive Interview)*. Machine Learning Street Talk (YouTube). Interview with Matej Balog and Alexander Novikov on AlphaEvolve. URL: https://www.youtube.com/watch?v=vC9nAosXrJw.

Moskvichev, Arseny, Victor Vikram Odouard, and Melanie Mitchell (2023). "The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain". In: *arXiv preprint arXiv:2305.07141*. arXiv: 2305.07141 [cs.AI]. URL: https://arxiv.org/abs/2305.07141.

Nguyen, Timothy (2024a). "Understanding Transformers via N-gram Statistics". In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/b1c446eebd9a317dd0e96b16908c821a-Paper-Conference.pdf.

— (2024b). *Understanding Transformers via N-Gram Statistics: Timothy Nguyen*. Machine Learning Street Talk (YouTube). Interview on transformer mechanics and n-gram statistics. URL: https://www.youtube.com/watch?v=W485bz0_TdI.

Novikov, Alexander, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog (2025). "AlphaEvolve: A coding agent for scientific and algorithmic discovery". In: *Google DeepMind Technical Report*. Describes the AlphaEvolve system for evolutionary refinement of LLM-generated code. URL: https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/AlphaEvolve.pdf.

Risi, Sebastian, Yujin Tang, David Ha, and Risto Miikkulainen (2025). *Neuroevolution: Harnessing Creativity in AI Agent Design*. Cambridge, MA: MIT Press. URL: https://neuroevolutionbook.com.

Rocktäschel, Tim (2024). *Open-Ended AI: The Key to Superhuman Intelligence? – Prof. Tim Rocktäschel*. Machine Learning Street Talk (YouTube). Interview on open-endedness, creativity, and the formal definition of open-ended systems. URL: https://www.youtube.com/watch?v=6DrCq8Ry2cw.

Romera-Paredes, Bernardino, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi (2024). "Mathematical discoveries from program search with large language models". In: *Nature* 625. Introduces FunSearch, a precursor to AlphaEvolve, pp. 468–475. DOI: 10.1038/s41586-023-06924-6. URL: https://doi.org/10.1038/s41586-023-06924-6.

Runco, Mark A. (2023). "Updating the Standard Definition of Creativity to Account for the Artificial Creativity of AI". In: *Creativity Research Journal* 37.1, pp. 1–5. DOI: 10.1080/10400419.2023.2257977. URL: https://www.tandfonline.com/doi/abs/10.1080/10400419.2023.2257977.

Runco, Mark A. and Garrett J. Jaeger (2012). "The Standard Definition of Creativity". In: *Creativity Research Journal* 24.1, pp. 92–96. DOI: 10.1080/10400419.2012.650092. URL: https://www.tandfonline.com/doi/abs/10.1080/10400419.2012.650092.

Schlosser, Markus (2019). "Agency". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Comprehensive overview of philosophical theories of agency, from minimal agency in simple organisms to full rational autonomy. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/entries/agency/.

Shin, Minkyu, Jin Kim, Bas van Opheusden, and Thomas L. Griffiths (2023). "Superhuman artificial intelligence can improve human decision-making by increasing novelty". In: *Proceedings of the National Academy of Sciences* 120.12, e2214840120. DOI: 10.1073/pnas.2214840120. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2214840120.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal (2024). "AI models collapse when trained on recursively generated data". In: *Nature* 631, pp. 755–759. DOI: 10.1038/s41586-024-07566-y. URL: https://www.nature.com/articles/s41586-024-07566-y.

Silver, David, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis (2017). "Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm". In: *arXiv preprint arXiv:1712.01815*. arXiv: 1712.01815 [cs.AI]. URL: https://arxiv.org/abs/1712.01815.

— (2018). "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419, pp. 1140–1144. DOI: 10.1126/science.aar6404. URL: https://www.science.org/doi/10.1126/science.aar6404.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis (2017). "Mastering the game of Go without human knowledge". In: *Nature* 550.7676, pp. 354–359. DOI: 10.1038/nature24270. URL: https://doi.org/10.1038/nature24270.

Sims, Karl (1991). "Artificial Evolution for Computer Graphics". In: *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '91)*. ACM, pp. 319–328. DOI: 10.1145/122718.122752. URL: https://dl.acm.org/doi/abs/10.1145/122718.122752.

Stanley, Kenneth O. (2014). *Innovation Workshop: Open-Ended Discovery of Ideas*. Santa Fe Institute Workshop. Presentation on Picbreeder and open-ended evolution. URL: https://wiki.santafe.edu/images/3/34/Stanley_innovation_workshop14.pdf.

— (2021). *Kenneth Stanley: Abandoning Objectives for AI Innovation*. Machine Learning Street Talk (YouTube). Interview on open-endedness, novelty search, and AI creativity. URL: https://www.youtube.com/watch?v=lhYGXYeMq_E.

— (2025). *Creativity is the ability to make intelligent decisions without a destination in mind.* Twitter/X. Tweet on creativity and LLM limitations. URL: https://x.com/kenneth0stanley/status/1931423482942017688.

Stanley, Kenneth O. and Akarsh Kumar (2025a). *AI is SO Smart, Why Are Its Internals 'Spaghetti'?* Machine Learning Street Talk (YouTube). Interview on the FER paper: fractured and entangled representations vs open-ended discovery. URL: https://www.youtube.com/watch?v=o1q6Hhz0MAg.

— (2025b). *Kenneth Stanley: The Power of Open-Ended Search Representations*. Machine Learning Street Talk (YouTube). Interview on open-ended search, derivative vs transformative creativity, and representations. URL: https://www.youtube.com/watch?v=KKUKikuV58o.

Stanley, Kenneth O. and Joel Lehman (2015). *Why Greatness Cannot Be Planned: The Myth of the Objective*. Cham, Switzerland: Springer. ISBN: 978-3319155234. DOI: 10.1007/978-3-319-15524-1. URL: https://doi.org/10.1007/978-3-319-15524-1.

Tian, Yufei, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman (2024). "MacGyver: Are Large Language Models Creative Problem Solvers?" In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*. Introduces the MacGyver benchmark for creative problem solving. arXiv: 2311.09682 [cs.CL]. URL: https://arxiv.org/abs/2311.09682.

Turing, Alan M. (1950). "Computing Machinery and Intelligence". In: *Mind* 59.236, pp. 433–460. DOI: 10.1093/mind/LIX.236.433. URL: https://doi.org/10.1093/mind/LIX.236.433.

Wang, Tony T., Adam Gleave, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell (2023). "Adversarial Policies Beat Superhuman Go AIs". In: *arXiv preprint arXiv:2211.00241*. Demonstrates adversarial strategies that exploit weaknesses in superhuman Go AIs. arXiv: 2211.00241 [cs.LG]. URL: https://arxiv.org/abs/2211.00241.

How to Cite This Article

Budd, J. & Scarfe, T. (2026). Why Greatness Cannot Be Interpolated. *MLST Archive*. https://archive.mlst.ai/paper/why-greatness-cannot-be-interpolated

BibTeX:

```
@article{mlst_2026_001,
  title   = {Why Greatness Cannot Be Interpolated},
  author  = {Jeremy Budd and Tim Scarfe},
  journal = {MLST Archive},
  year    = {2026},
  url     = {https://archive.mlst.ai/paper/
             why-greatness-cannot-be-interpolated}
}
```