CONJECTURE INSTITUTE · FELLOW SPOTLIGHT

The Three Pillars of an Artificial Mind

The road to artificial intelligence (AG) is not paved in increasingly efficient algorithms, more computing power, or radical improvements to the best large language models. Armed with our best theory of how knowledge is created, Carlos has honed in on universality, problem-solving, and explanatory knowledge as three key pillars to building an AGI.

BY LOGAN CHIPKIN

Image by Amaro Koberle

Progress in Mind

"I read David Deutsch's 2012 essay, *Creative Blocks*, about how there might just be a single idea standing between us and the recipe for building an artificial general intelligence (AGI)," Carlos recalls. "Deutsch wrote that it would be one of the best ideas that anyone's ever had. And I thought, 'Somebody should do something about that. Well, I'll give it a go.'"

Carlos had finished his undergraduate program in mechanical engineering that same year, educated in parallel on YouTube by economist Milton Friedman, polemical journalist Christopher Hitchens, and New Atheist Sam Harris. After seeing Deutsch's book, *The Beginning of Infinity*, on Harris' book list, Carlos picked it up.

It would forever alter the course of his career.

First, he breathed in the book's expansive, cosmic optimism, as well as its explanation of twentieth-century philosopher Karl Popper's theory of knowledge. Contrary to the prevailing notions of empiricism—the theory that knowledge derives from the senses—and inductivism—the theory that we induce our way to theories of the world by making repeated observations—Popper's theory of knowledge said that all knowledge grows by first boldly conjecturing one or more theories that solve a problem, criticizing all of them, and then tentatively retaining whichever survives. Armed with the vocabulary of conjectures and refutations, he'd puzzle out how to recast every un-Popperian statement he'd hear in Popperian terms.

Why Care about Epistemology?

WATCH ON YOUTUBE

Why care about knowledge and knowledge creation in the first place? Put simply, everything that we care about and everything that's interesting depends on it—from a person's daily life to whether an institution succeeds or fails to the dynamics of an entire economy or civilization. Indeed, the human brain is special precisely because of its ability to create explanatory knowledge—accounts of how the world works and why.

In biology, knowledge is a predominant factor. The emergence of new mutations and species corresponds to the birth of new genetic knowledge.

Aliens must largely consist of knowledge. Whatever genetic code caused them to come about in the first place, and whatever technology they might create, both depend on knowledge.

An artificial general intelligence's abilities depend on the knowledge they *have* created and *can* create, and whether they're dangerous depends on how good they are at creating knowledge relative to us.

Individual animals learn things during their lives, and we can ask how their methods of knowledge creation compare to human methods. It seems that humans are exceptional—no animal is getting to the moon except humans, and that demands explanation.

Carlos groups all of these things—animals, genes, species, individuals, groups of people, institutions, aliens, artificial general intelligence—into a single set called 'knowledge-creating systems'. Despite their differences, despite the fact that daily life and its concerns are quite different from the concerns of how humans work in general or groups of humans or biological systems or biological evolution as a whole or aliens or AGI or animals, all of these things *use* and, more importantly, *create* knowledge.

This unification means that all of the fundamental insights that we can glean about knowledge help to explain the most important features of all of these systems.

And when it comes to knowledge, there are only a few basic questions we need to concern ourselves with. One question we can ask of any one of these systems—again, whether we're talking about a human mind, an institution, genes, an alien, an animal's brain, or an AI program—is: what kind of effects do its bits of knowledge have on other bits of knowledge or on other physical systems? A second question we can ask is not about the knowledge's *effects* but rather its *causes*—how did the knowledge come to be? What sequence of events caused the knowledge to come about? Which parts of its history were merely *contingent*, and which were *necessary*?

Now, for any knowledge-creating system to improve, it must be able to do two things. First, it has to be able to *generate* new ideas, actions, strategies, etc. Knowledge-creation can't get off the ground if the system in question is incapable of generating new options or else creatively varying its past suite of options so that it can try out new ones. To discover *better* ideas, genes, explanations, institutional norms, etc., you first need to create *new* ones—not all of which will be improvements.

The second key process is *condensing* or *filtering*—all knowledge-creating systems need some kind of filter that can take every new idea and retain the good ones and kill the bad ones.

So we have this pair of processes: generation or variation, this chaotic creation of new things, some of which are good, some of which are bad. And then we have selection or filtering, the identification of the good things among the set of things we've created and the erasure of the bad things we've created.

We can compare various knowledge-creating systems in light of what we've just discussed. For example, what are the limits on the kinds of knowledge that an animal can create compared with a person? A democratic polity compared with a monarchy?

Even if two knowledge-creating systems can in principle generate the same set of knowledge, might one be more efficient than another? That is, might one's filtering system be better than another's?

Similarly, even if two knowledge-creating systems can in principle create the same set of ideas, might one's *generating* process be more prolific than another's?

Carlos' first job (and firing) out of college was in robotics. After pushing a passage from *The Beginning of Infinity* on a superior, the reply came back: "Carlos, why can't you be this excited about *work*?"

"I took the advice, in a way," Carlos says. "I took those exciting ideas, and made them my work. After some detours in data science and longevity research, I've settled into the delightful chaos that is open-ended research, all revolving around how minds and progress are made."

Carlos left his robotics job without a plan, but he knew that he never wanted to commit to something he didn't enjoy again. He moved back to College Station and Texas A&M, where he still had

an apartment lease, and quickly fell into his old habit of devouring books and videos about fundamental ideas, finally deciding to read Karl Popper directly. It was during this time that Carlos read Deutsch's article about AGI.

By the time his lease ended, Carlos was well-versed in the ideas of epistemology, progress, and AGI, but he still didn't see himself as anything like an AGI researcher.

"I was a little aimless around then," Carlos says.

He moved to New York City for his next adventure, where he dabbled in building apps, coding, and he even landed a Silicon Valley job in biotech at one point. His interest in epistemology simmered, helped by a growing group of Deutsch's fans on Twitter.

His biotech gig would end like his robotics job did (though without the firing), but this time Carlos had more savings, more knowledge, and more self-assuredness.

Almost without realizing it, he became a 'solopreneur', an entrepreneur who sells ideas and products directly to customers without any kind of supporting staff.

"My interest in researching and writing about AGI didn't hit me at once," Carlos says. "In the years after my undergrad, it was a slow burn that grew over time. No matter what else I was doing, the ideas never quite left me."

As a solopreneur, Carlos tried (and succeeded at) a bunch of different projects: writing essays on Substack, hosting workshops, fielding private one-on-one lessons, and creating lightboard videos. By now, his interests were well-developed, and all of his projects were colored by his latest ideas on epistemology, progress, and AGI.

The Five Nested Sets of the Mind

Carlos has discovered a useful categorization for the mind's ideas: what he calls the *nested sets of the mind*. In nested order, these are: Active ⊆ Existing ⊆ N-Reachable ⊆ ω-Reachable ⊆ Possible.

1. ***Active***. The set of things currently active in the mind. This is usually a tiny set, focused only on what is relevant at any given moment. The fact that we can focus like this helps us solve problems more efficiently than just trying things at random.

2. ***Existing***. All the knowledge in the mind. Many of these things have not been used in years, but they are in there, and might be used if they suddenly become relevant. Ideally, you would have vast knowledge, and know exactly when to use each thing—when it should become active.

3. ***N-Reachable***. All the knowledge you do not yet have, but might reach in N steps. Many things are just a few thoughts away. Also, if your mind contains good ideas, then many improvements might be just a few steps away.

4. ***ω-Reachable***. All the things you can ever reach, given unlimited time and resources. The human mind might be able to create any thought, while AIs or raccoons might never reach those things, even with infinite time.

5. ***Possible***. All the things that are possible in reality. Some systems can do all these things—perhaps we humans can. If the set is finite, then there is a best possible state, and things can never improve after that. If the set is infinite, then perhaps things can get better forever.

Implicit in Carlos' framework is that an artificial general intelligence must, at any given moment, have sets of active, existing, and N-Reachable ideas that can be characterized *objectively*, without reference to subjective measures, probabilities, or weights. Moreover, regardless of the size of these sets, an artificial general intelligence must always be able to progress arbitrarily towards ω-Reachable ideas.

Carlos' Three Pillars of AGI

Carlos pursues artificial general intelligence from a wholly Popperian perspective: rather than asking how much computing power is required to build one, or asking how an artificial general intelligence needs to be programmed to update its hypotheses in light of new data, Carlos asks questions such as: What ideas can a mind possibly think? What does it take for a mind to make progress from a worse set of ideas to a better one, and do it efficiently?

Armed with Popperian epistemology and related insights from Deutsch's *The Beginning of Infinity*, Carlos has honed in on *universality*, *problem-solving*, and *explanatory knowledge* as three key pillars to building an AGI.

"So my thinking was," Carlos recalls from his early days of researching AGI, "nobody's taken all three of these things seriously. What if somebody takes these for all they're worth and just dives into them? And I've been thinking about these three pillars for years since then."

The universality of the human mind, or of creative thinking, is a kind of demarcator—either you have it or you don't.

"Most people imagine that intelligence is a scale," Carlos says. "Researchers mistakenly think that you can have a little, you can have a lot, it can keep on going. They assume that humans are in the middle somewhere, and that you can imagine something becoming a little smarter than us and maybe much smarter—as in the supposed 'superintelligence'."

"The human brain does have Turing completeness, which is basically the idea that it can run any program. But this isn't enough for intelligence, as it's ubiquitous. Still, I think the idea of Turing completeness is a sort of existence proof that there can be a binary

distinction that is all-important. I think the difference between humans and other animals might be like that, perhaps even built on that."

In short, a universality of some kind, as yet poorly understood, might be the key distinction between human-level intelligence as one type of knowledge-creating system and literally everything else that is lesser than that.

Five Fundamental Aspects of Creative Thought

WATCH ON YOUTUBE

Another insight that has come out of Carlos' research is the five dimensions along which one may compare humans' and artificial general intelligence's ability to generate ideas. Are there any limits on knowledge creation, on how thinking works, that bound humans but not artificial general intelligence?

Carlos approaches the question by considering five fundamental aspects of creative thinking:

1. ***Computation***. Is there a limit to what we can compute? Not once we have a universal computer, which we do. Provided a person has enough time and memory, they can compute anything that can possibly be computed. AGIs would have precisely this same repertoire.

2. ***Transformations***. It seems that humans are capable of causing any transformation of raw materials into a final product that is allowed by the laws of physics. Any given transformation can be rendered by executing a sequence of rudimentary steps, all of which a human is capable of (and the more advanced our technology, the wider the set of transformations that we can execute at that point in time). Human hands can build a tool that can build a tool that can build a final product. An AGI may start off with something different than human hands, but they could converge on humans' sequence of tools in short order—and vice versa.

3. ***Variation***. Do humans face any limitation in the kinds of new ideas that they can generate? No, if only because we can already navigate the space of all possible computer programs by exhaustive, brute force navigation. But humans can vary their ideas exponentially more efficiently than that—not only can we navigate the entire space of ideas, but we can navigate it in any physically possible way, some far more efficient than others. AGI will also have this capability, and so neither humans nor AGI will be privileged with respect to their ability to vary their ideas.

4. ***Selection***. Coming up with new ideas is worthless if we cannot select between better and worse ideas. Cumulative progress requires multiple cycles (of arbitrary number) of selecting better ideas over worse ones. When you come up with new options, you need to be able to compare them according to some criteria. It seems that humans are capable of improving these criteria themselves ad infinitum. Therefore, humans' ability to select between ideas is itself capable of infinite improvement, which means that AGIs cannot supersede us with respect to selection.

5. ***Attention***. It is little good having lots of ideas if you cannot apply the right ones in the right problem situations. If you stub your toe, it is far better to think about how to move your toe away from the thing that you hit rather than what crocodiles eat to survive, even though you happen to also have ideas about the latter. Randomly cycling through all of the ideas in your mind in an effort to solve the problem situation in which you find yourself would result in surefire death. If we create AGIs with better ability to direct their attention, then humans can creatively improve upon our internal algorithms and catch up to them.

The second pillar, the Popperian idea of problem-solving, stands in violent contrast to the empiricism (the false idea that knowledge derives from the senses) that dominates AGI research. To reiterate the Popperian scheme of things: you face a problem, you come up with some ideas that might solve it, you criticize all of the candidate solutions, and you tentatively adopt whichever of them (if any) has no outstanding criticisms against it. That leaves you in a better situation, at which point the entire process begins anew.

"That's the core engine," Carlos says. "Again, this is in contrast to the popular idea that you get a bunch of data, you get the patterns, and you somehow get knowledge out of those."

The third pillar, explanatory knowledge, is a special kind of knowledge that, apparently, only humans and other human-level intelligences (like advanced aliens and AGIs) are capable of producing. Explanatory knowledge is an account of how the world works and why. Carlos keeps the distinction between explanatory and non-explanatory knowledge in mind (which itself comes with many open questions) and thinks about how this fits in with the other two pillars to form a person-type mind.

"So that was my conception of things as I began my research," Carlos recalls. "From there, it's been a matter of sharpening my understanding of each of the pillars, as well as how they might

relate to one another."

For example, with respect to universality, one of Carlos' lines of research is trying to understand the different types of universality and whether any of them are adequate for building an AGI, or if some are not up to the task.

With respect to Popperian problem-solving, Carlos noticed that, although Popper's scheme has a lot in common with Darwin's scheme for biological evolution, the latter has no conception of problems.

"Darwin's motto is variation and selection," Carlos says. "That's it. But Popper added a third thing: problems. How important is that? Is that *necessary* for progress, or is it just a 'nice to have'? What's going on there? I eventually concluded that the presence of problems is a way of accelerating progress."

If you can know what to focus on, which the concept of a problem necessarily brings along for the ride, then you can make all of your variation and selection tailored to a very narrow band of possibilities. Biological evolution lacks this, which is one reason—maybe the primary reason—why it is so much slower than the progress that people make.

The other aspect of problem-solving that Carlos focused on was Deutsch's description of problems in *The Beginning of Infinity* as 'conflicts between ideas'.

"That seemed a bit awkward to me," Carlos says, "because the more general thing is not merely conflicts, but just anything that makes you focus toward or away from certain things. So even if there's merely a very vague sense of, 'this is cool' or 'this is bad', those sorts of thoughts are still extremely helpful in terms of focusing your creative energies."

Carlos reasoned that more general than a 'conflict between ideas' is the notion of *attention*—choosing to focus your resources *here* rather than *there*. Moreover, the vast majority of your ideas and options at any given moment are irrelevant to the situation at hand and are very unlikely to lead anywhere useful—thinking about dolphins will not help you with your construction job. Attention is a necessary element in making rapid progress and knowledge growth, which any AGI must be capable of.

"I've developed a more concrete view of the second pillar, of my broad conception of Popperian problem-solving. Progress and creative thought are not only about *coming up* with new things and comparing them, but they also entail having a way of *choosing*, of using any possible thing available, any heuristic, that helps you determine whether to focus your energies here or there."

Unlike the overwhelming majority of other artificial general intelligence researchers, Carlos is not trying to transform current artificial intelligence into a human-like entity, nor is he trying to algorithmize intelligence. He is using our best epistemology to distinguish the creative mind from uncreative matter.

Carlos' Four Requirements for Sustained Progress

1. ***Novelty***. If you try new things, there's no guarantee they'll be better than what you started with. However, if you never try anything new, you are *guaranteed* to never improve yourself. Novelty just means doing something new. That said, there are different measures of novelty. A box of bouncing molecules is always doing something new—you will never see the molecules in the same configuration twice. But if you ask what it looks like, it always looks random. In that sense, it never does anything new. There are infinitely many ways to look at the box and what it is doing, infinitely many measures of novelty.

2. ***Taste***. Like measures of novelty, a system needs measures of quality. Which animal is fastest or strongest, which machine learning model is best at face recognition, which recipe tastes best, which scientific theory makes the most sense. Without such measures, a system is exploring the possibilities at random, unlikely to find good stuff, and especially unlikely to recognize good stuff and focus on it.

3. ***Control***. It is one thing to have a sense of quality—what's good and what's bad, and in what way. But, you have to be able to *do something* with that information. If you couldn't, then you'd be like a consultant at a company that nobody listened to. You might know what to do, but that will have zero impact on what happens. If the company is making a mistake, then they'll go on making it.

4. ***Memory***. To make continued progress, you need not only to *make* improvements, but *keep* them, so you can build on them. If you have amnesia and your library burns down every day, then you have to start from scratch, and can never discover or create anything that requires more than a day of work. If you want to invest centuries into a cathedral, or a billion years into ever more complex living organisms, you'll need to save your work along the way. The bricks of the cathedral need to stay where you put them. The genes that specify an organism need to persist—either in one instance, or in copies. Otherwise, evolution would have to start from scratch—and giraffes can't be built in a day.

Without novelty, things remain the same. Ask: what are the sources of novelty in the system? What kinds of novelty?

Without taste, the good and the bad are indistinguishable. Ask: how does the system distinguish the good from the bad? What measures does it use? Can these change over time?

Without control, taste has no impact. Ask: if a system has taste, how does that impact what happens? Once something is determined to be good or bad, what is done with that information?

Without memory, all prior work is lost. Ask: when something happens in the system, how does it affect the distant future of the system? How are the results saved? Where is the information stored?

Popperian Engines and Knowledge Goggles

Lately, Carlos has been looking at AGI from two opposite directions. On the one hand, he tries to break down his three pillars into ever smaller subunits, much as particle physicists try to learn about how the world works by analyzing the dynamics of the smallest possible units of reality.

On the other hand, Carlos investigates how one can integrate the three pillars into a unified whole. What are the simplest mechanisms or processes by which universality, problem-solving, and the ability to generate explanatory knowledge can become embedded in a single object (i.e. a mind)?

In parallel to his research, Carlos is also working on a book about AGI, though it is sure to be as much about epistemology and the nature of progress as it will be about Carlos' insights into how we might create an AGI from scratch.

"To me, the engine of AGI is clearly Popperian epistemology," Carlos says. "So to get a better picture of AGI, you have to sharpen your fundamental tools when it comes to epistemology: the concepts of problems, ideas, criticism, and attention can't be ignored. And epistemology is so broad that, even though I'm focused on AGI, I can't help but be occasionally drawn to other fields where the growth of knowledge is important, like biology and economics."

Carlos sees the world—not just AGI, but the entire cosmos—through the knowledge-based view of things. Whereas most people look out at the night sky and see patches of darkness interspersed with distant points of light, Carlos regards knowledge-creating entities as the true 'lights' of the universe, and everything else as effectively 'dark'.

"Imagine putting on goggles that only let you see resilient patterns of information—knowledge in other words," Carlos invites us to try. "When you look at stars, suddenly they're dark. But when you look at crabs, dogs, and even trees, they radiate some light from every strand of DNA in every cell. And when your eyes land on *people*, you see the same glow from the genes, but also a new and far more intense glow from their brains."

The vast majority of other AGI researchers do not wear these goggles. They insist that if we somehow just add more 'dark' things to the mix—more data, more efficient algorithms, better hardware—the light that is a knowledge-creating person will somehow mechanically pop out. But Carlos is among the few who understands that our best ideas about personhood, knowledge-creation, and computation imply a radically different path to AGI.

Armed with his three pillars and the knowledge-based view, he's doing something about it.

Originally published by Conjecture Institute at conjectureinstitute.org/articles/fellow-spotlight/carlos-de-la-guardia.