Conjecture Institute Handbook

CONJECTURE INSTITUTE

Handbook

A Summary of Our Mission,
Worldview, Projects, and People

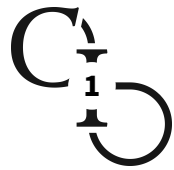
PUBLISHED MARCH 2026

First published December 2025

FIRST EDITION

Conjecture Institute
https://conjectureinstitute.org

Contents



Having the Right Ideas Matters

Epistemology: The Wrong Questions

The history of political philosophy is littered with questions such as: Who has the right to rule? Which political structure can best ensure that the right rulers rise to power? Is democracy, typically understood as 'the will of the people', really superior to a monarchical order that manages to keep competent rulers on the throne?

Arguments over policies tend towards similarly fallacious reasoning, such as: What evidence is there that such-and-such policy works? What data justifies this or that policy?

The quest for artificial general intelligence is similarly trodden by researchers on roads to nowhere. Some think that it is a matter of crossing some resource threshold: once we have an LLM that can consume a particular amount of energy per unit time, or can digest some amount of bits per second, then we will have discovered artificial general intelligence. Still others think that it's a matter of finding the most efficient pattern-recognition algorithm. Some think that it's a matter of emulating (or approximately emulating) the input-output functions that take place in the human brain and then grafting these functions onto silicon.

Some thinkers are convinced that artificial general intelligence will be some kind of induction machine, one that will take an arbitrary batch of data and induce ideas from that. Others think that it will be a probabilistic machine, one that continuously updates the likelihood of various hypotheses in light of new data.

Related to these roads to nowhere are worries of doomsday scenarios: if we create an artificial general intelligence that can think faster than humans, so the argument goes, then it will quickly outstrip us and dominate us just as we dominate ants. Another worry is that artificial intelligence will destroy the world in a kind of 'paperclip maximizing' event: if programmed with a value such as 'create the most amount of paperclips possible', the technology will endeavor to transform as much of the universe as it can into paperclips.

The field of economics is riddled with esoteric mathematical models, as what's known as 'physics envy' has consumed entire swathes of this fundamental discipline. They insist that nations, companies, and individuals correspond to some mathematical abstractions.

The hard sciences fare little better, as they suffer from a number of unforced errors, such as instrumentalism, anti-realism, empiricism, and fetishization of mathematics.

Many physicists in particular are convinced that the next fundamental theory must be in terms of ever smaller particles (or strings, as the case may be). They similarly think that any experimental breakthroughs must come from bigger and bigger particle colliders that will answer something like: What are the universe's smallest, most fundamental building blocks?

Physicists also tend to cast physical problems in purely mathematical terms. For example, in trying to reconcile quantum mechanics and general relativity, they seek to answer: What general mathematical object might contain the machinery of both theories as limiting cases?

There is an entire industry of how-to books for parenting in the modern age. They give comfort to parents who worry over questions such as: How do I make sure my kid gets into the best college? How do I keep my child grounded in a world full of distractions? How can I keep my child away from the addictive vices of screens and social media? How can I discipline my child without resorting to the barbarity of spanking?

In philosophy, intellectuals continue to spend their careers on questions such as: How are our scientific theories justified? How can we be certain about anything? Which moral theory—whether consequentialism, deontology, or anything in between—is correct, and how do we know?

Some of these fields—politics, artificial general intelligence, economics, and parenting—are about people and their ideas. Therefore, our understanding of what knowledge is and how it can be improved upon (if at all) shapes the very questions we ask in all of these domains. The content of our ideas in political philosophy and parenting philosophy is intertwined with the nature of

knowledge, how we acquire it, and how it can grow. Even physics, which is not primarily about people but rather the workings of the universe, is a human enterprise whose progress or lack thereof is inextricably tied up with how we treat the nature of knowledge (this is the so-called philosophy of science). So while the *content* of our ideas in physics may not refer to our philosophy of ideas, the *process* by which we 'do' physics very much does—even if only implicitly.

So there is no getting around our conception of what knowledge is and how it grows. Philosophy matters, after all.

Epistemology: The Right Questions

Our deepest theory of knowledge was discovered by the 20th century philosopher, Karl Popper. In a sentence, it states that *knowledge grows via conjecture and criticism*.

As with many of our other fundamental theories of reality, the central thesis of Popper's theory of knowledge—also known as *critical rationalism*—has enormous and counterintuitive ramifications for our worldview. It rules out as nonsensical any quest for increasing the probability of our theories being true, our degree of confidence in our hypotheses, and any criterion for either truth or certainty. It implies that the growth of knowledge is unpredictable, and so mechanistic models of any system in which knowledge plays a role are approximative at best and grossly misleading at worst. It contradicts *empiricism*, the idea that knowledge is derived from the senses, as well as *inductivism*, the idea that the past will resemble the future. And finally, it suggests that all of our ideas are fallible—devoid of certainty, riddled with errors, forever subject to improvement.

Physicist and Conjecture Institute Advisor David Deutsch added an indispensable epistemological principle to our collective arsenal: *all progress consists of the quest for good explanations—accounts of the world that are hard to vary*.

A good explanation—whether in science, politics, morality, or any other domain—is one whose ability to account for what it purports to account for would collapse if any of its components were changed. If you replace even one conceptual or mathematical element of classical mechanics, for example, the entire explanation loses its coherence. Replace acceleration with velocity in Newton's Second Law, then Newton's First Law wouldn't work, because then objects ought to slow to a halt in the absence of external forces. One can play with the elements of the theory in this way, permuting them as one wishes, only to find

that most permutations would render other parts of the theory problematic (to say nothing of the disintegration of the theory's predictive powers). Newton's 'version' of the theory as he presented it is coherent, and delicately so—it is hard to vary while retaining its ability to explain (and accurately predict) the dynamics of massive objects.

This principle of seeking good explanations can be fused with critical rationalism's central tenet: *all knowledge grows by guessing and criticizing good explanations.*

With this upgraded critical rationalism in hand, we can rule out whole alleys of what are, with the benefit of hindsight, paths that lead only to ditches, wasted effort, and counterproductive choices.

In political philosophy, the right question is not, "Who should rule?" since everyone is fallible. Rather, since any given ruler's outlook will contain errors, a polity needs a process that allows for the correction of those errors by way of nonviolent replacement. The better questions, then, are "What sort of political institutions allow for the peaceful transition of power?" and "How can we optimize a political institution for error correction?" and "How can we arrange our political system so that politicians are held responsible for the results of their policies?" This Popperian analysis clearly favors democracy over monarchy and dictatorship. And even between democracies, it favors systems that allow for politicians to implement their policies unblemished by compromise over those that pressure politicians to compromise amongst each other. The implementation of a policy is a kind of test (read: criticism) of that policy and its underlying political worldview (read: good explanation), and it makes as much sense to mix various policies together as it does to mix scientific theories together (read: mixing two good explanations together does not necessarily yield a good explanation).

The fact that knowledge grows via conjecturing good explanations and criticizing them immediately rules out any pursuit of artificial general intelligence that rely on empiricism or inductivism. Contrary to the popular wisdom of empiricism, a mind does not blindly absorb ideas via sensory data—one does not derive the physics of the solar system upon receiving light from the sky, nor does one derive biology from visiting one thousand or one million different species. On the contrary, a mind first *guesses* a good explanation for the motion of the objects in the sky or the intricate designs of the bird's wing and the lion's claw. Only *then* does data come into the picture, as a *criticism* of our guesses.

What would hardcoding empiricism even look like? Maybe the software engineer designs a general formula such that the artificial intelligence attributes a

particular hypothesis to a particular data set and updates according to some fixed set of rules. But the source of the conjecture is the *engineer*, not the *artificial intelligence*. A true artificial general intelligence would, like a person, be capable of producing an endless stream of conjectures that do *not* depend on experiential data nor hardcoded criteria.

Hardcoding inductivism into software can likewise never produce an artificial general intelligence. In some ways, the future will resemble the past, and future observations will resemble past observations. But deciding *which* aspects of reality will repeat themselves (and under what conditions) is a conjectural project, not one that is either set in stone or available to us a priori. An artificial intelligence might well be able to detect patterns in data, but *which* patterns it detects are ultimately attributable to immutable code that a software engineer had embedded into the technology, and that code cannot itself be indefinitely improved upon by the artificial intelligence. A genuine artificial general intelligence would be able to guess regularities in Nature that no one had ever previously considered, proffer up patterns that satisfy novel criteria, guess ideas about the world that could *not* be traced back to its code.

No matter how sophisticated the outputs of the latest artificial intelligence are, they are not fundamentally new in a very specific sense: all of their outputs can be explained by the algorithm that constitutes it. All of their supposed conjectures, pattern-recognition, and data gathering techniques are mechanically determined by their software. They are slaves to the closed system that their software engineer had created. An artificial *general* intelligence, on the other hand, is slave to nothing but the laws of physics. *His* conjectures cannot be explained by examining the algorithm that constitutes his mind (to say nothing of his genetic code). The good explanations that he comes up with are genuinely new things in the world, derivable from nothing. Whereas the software engineer is the true father of an artificial intelligence's output, a person is the father of his own thoughts. We will know that we have created an artificial general intelligence when its creative output cannot be explained by the workings of its software engineer but rather by its own creative thought.

The dichotomy between artificial intelligence's closed, slavish character and a person's open-ended, creative character sheds light on the doomer scenarios briefly described above. For if an artificial intelligence attempts to turn the universe into paperclips, it can only do so with the ideas that had been coded into it by an engineer. All of civilization, meanwhile, stands ready to come up with new ideas with which to halt the paperclip maximizer against which the automaton has no defense.

Doomsday scenarios that envision a dark artificial general intelligence that will think orders of magnitude faster than us and come to dominate us by way of its newfound knowledge simply assume that a fast-thinking entity would 1) have knowledge completely divorced from Western institutions, and 2) not make egregious mistakes on its way to godhood. An artificial general intelligence is a person and, no matter how fast it thinks, it must first *catch up* to the knowledge of the culture into which it is born before outpacing it. Both activities are conjectural, and the West does an amazing job at assimilating its newborns into a peaceful, cooperative society. Should the artificial general intelligence still choose to become a criminal or a terrorist, outpacing the West to such a degree that no one can fight back is *still* a conjectural affair, and one that it can get catastrophically wrong—for itself. To give up on the quest for artificial general intelligence because it might want to destroy the world is to give up on creating more human babies. After all, they too might grow up to want to destroy the world. Because all of our ideas are fallible, nothing is guaranteed. But an artificial general intelligence that mistakenly wants to ruin society is no infallible enemy—like any other enemy of civilization, we could fight back if we so choose.

Some doomsday scenarios picture an all-powerful artificial general intelligence whose values are forever orthogonal to our own. But if the silicon-based entity is truly as creative as a person, then it cannot have fixed values. Values, like all ideas, are forever fallible, conjectural, and subject to improvement. Just as individuals from vastly different cultures are capable of converging on values, goals, and scientific theories, so it is with artificial general intelligence and the rest of humanity. This sort of doomsday scenario that imagines an incorrigibly dangerous artificial general intelligence is therefore as rational a consideration as worrying that birds might give birth to bloodthirsty dinosaurs—a physical impossibility.

The fact that all knowledge is creatively guessed immediately rules out any and all deterministic models of economic activity. The 'physics envy' that plagues so many economics departments is misplaced—unlike a collection of particles or planets, the economy cannot be captured by a set of algebraic equations. There can be no equation that quantitatively predicts how a given good's price changes following a change in the supply of or demand for that good. Nor can a relationship between algebraic symbols ever correspond to the relationship between, say, interest rates and purchasing power. Those economists who insist that algebraic, deterministic models must play a role in explaining the regularities that constitute an economy are implicitly insisting on a false epistemology—namely that people

are predictable, uncreative automata. So much for econophysics and similar ways of trying to understand economics.

As we said in our discussion of artificial general intelligence, values are ideas—fallible, theoretical, and improvable. This is one reason that the labor theory of value is impossible—the economic value of a good is not derived from the work required to produce a good. Individuals *conjecture* the value of a good in light of their moral theories (including what they want), the tradeoffs they're willing and unwilling to make, and their knowledge of what they could and could not do with the good in question.

Finally, economic theories that implicitly assume the necessity of a particular institution do not make sense, since institutions are primarily constituted by the conjectured ideas of the individuals who operate them. If institutions are characterized by ideas, and if people can criticize those ideas and replace them with better ones, then no particular institution is necessary for an economy to exist. Therefore, any economic theory that assumes the existence of money, or a State, or even trade is at best incomplete.

So any fundamental theory of economics cannot be deterministic, it cannot assume the existence of any particular institution, and it must be consistent with critical rationalism. Econophysics and models that assume the necessity of the State are at best approximative, and the labor theory of value is ruled out from the start.

Austrian economics satisfies the aforementioned constraints—it conjectures that *individuals* act *purposefully* in a world of *scarce resources*. As with critical rationalism itself, this seemingly simple idea has wide-ranging and counterintuitive consequences for the nature of an economy. For example, one can deduce that socialism's problems are not just misaligned incentives (relative to capitalism), but that it faces fundamental limitations with respect to efficiently allocating resources.

In a monetary economy, consumers bid for various goods, the prices of which are determined by the consumers' demand for and the producers' (entrepreneurs or their associates) supply of them. These producers don't necessarily transform raw materials into their final consumer products: a restaurant owner might purchase utensils from a firm that specializes in the production of forks and knives. This firm, in turn, might buy the raw materials that are required in order to create utensils in the first place. For example, maybe they buy stainless steel from an owner of land who mines minerals solely in order to sell them. Physically, the production process of this example runs according to the following recipe: minerals are extracted from the earth by the landowner, then the fork-and-knife

firm transforms the minerals into utensils, then the restaurant owner 'transforms' them into a presentable meal, and finally the patron of the restaurant 'consumes' the presentable meal (part of which is the utensils). Economically, however, this entire production process flows *backwards* from the consumers' demand for consumer goods: the more patrons the restaurant owner serves, the more he, in turn, demands utensils from the fork-and-knife firm, which then demands more minerals from the landowner. This is what is meant when economists say that 'consumer is king'.

But the restaurant owner is not the only entrepreneur bidding for utensils—he is competing with homeowners, collectors, other restaurant owners, and speculators. Just as prices emerge on the consumer goods market between consumers and the entrepreneurs selling such goods, so too does a producer goods market between entrepreneurs and sellers of producer goods generate prices of *producer* goods. This is true for all stages of production, and again, this entire network of prices is ultimately driven by the consumers' demand for consumer goods.

When the entrepreneur sells a good to a consumer, the price of that consumer good is his *revenue*. In our example, this would be the restaurant owner selling a steak dinner to a patron for twenty dollars. But entrepreneurs do not pursue *revenue*, but rather *profit*. In order to calculate whether or not he's earned a profit, the restaurant owner must subtract the cost of producing the meal from the price at which he sold the meal. But what is that cost? It is the *price of the producer goods* that he purchased in order to create the meal in the first place—in our example, this includes not only the abovementioned utensils, but also whatever other producer goods went into the production of the meal. So, if the price of the sum of the producer goods exceeds the price of the final consumer good, the entrepreneur has incurred a loss. If the reverse is true, he's earned a profit.

Under voluntary conditions (a free market), profits and losses are *signals*—a profit indicates that the entrepreneur is satisfying consumers in transforming producer goods into consumer good that they demand. A loss indicates consumer dissatisfaction with that particular line of production. In either case, the entrepreneur may adjust his activities. In the first, he might demand more producer goods in order to try to earn even greater profits, while in the second, he might abandon his project in order to cut his losses. No matter what the entrepreneur does *next*, his decision will take in knowledge about what consumers want, and then transmit this knowledge 'backwards' to owners of producer goods. Without

the ability to calculate his profit/loss, the entrepreneur *cannot* know what to do next in order to better satisfy consumers.

Prices emerge on the producer goods market precisely because 1) these producer goods are privately owned, 2) entrepreneurs bid for them according to what they think consumers demand of *their* (the entrepreneurs') final consumer products, and 3) money is sound. Under centralized control of the means of production, *there is no producer goods market*. So the socialist institution has no idea what the prices of these producer goods are, and therefore has no profit/loss mechanism. Economic waste, such as shortages, surpluses, inefficient choices of which particular producer goods to employ in the creation of consumer goods, and reduction in total wealth (called economic regression) are all inevitable under a socialist order.

No matter how superhuman the "New Socialist Man" might be, no matter how selfless and communitarian, he will never overcome this calculation problem. The knowledge encoded in the profit/loss system that emerges in a free market cannot be recovered by a socialist Leviathan. Counterintuitively, it is centralized, coercive planning that causes destructive economic chaos, rather than an unplanned, voluntary system.

But is that really true? Had Mises shown that, following deductively from first principles, collective ownership of producer goods forces the socialist institution to provide consumer goods in wildly erroneous proportions relative to its decentralized, free market counterpart? The socialistic institution could always get *lucky*, and provide exactly what would've been provided in a free market. But then, a cow could similarly 'get lucky' and appear spontaneously in deep space. In neither case would we have a good explanation for why these seem to be regularities of nature.

A physical transformation is impossible if, no matter how much knowledge is brought to bear, it cannot be achieved. For example, building a generic spaceship out of raw materials is evidently possible, given that it has happened. However, building a particular spaceship that can travel faster than the speed of light is *impossible in principle*, since the laws of Einstein's special relativity forbid any massive object from traveling at such a speed. No matter how much more knowledge civilization acquires, we will never be able to violate the laws of nature.

The converse is also true—if no law of nature explicitly *forbids* a particular transformation from being achievable, then *people* are capable of causing said transformation, given the requisite knowledge. This implies that what we intuitively think of as *wealth* is more fundamentally about knowledge than about

the particular resources a person owns. For example, a farmer who owns the raw materials of land and seeds is capable of transforming them into edible crops, while a professor of history may not know what to do with those same resources. Furthermore, a person can grow wealthier without acquiring any new resources by instead acquiring more *knowledge*. The value of land with oil reserves only shot up in value *after* people learned how they could employ oil to their advantage, and not a moment before that. So the set of all possible transformations that the same scarce resources may undergo depends on what their owner *knows* what to do with them.

An economy is a particular way that knowledge is arranged in the universe, distributed across the minds of creative people. As we've seen, this knowledge can grow (or shrink, as when civilization regresses), causing a concomitant increase in economic wealth. Can we express this more exactly? Are there laws of nature that govern, constrain, and explain changes in the growth of knowledge and wealth?

Conjecture Institute is actively seeking a Fellow who will not only explain the principles and applications of Austrian economics to a lay audience, but who will also conduct original research into these questions. In the meantime, Logan is developing an economics course for Conjecture University.

Rules-based parenting and the school system implicitly appeal to the bucket theory of mind, the false idea that knowledge can be mechanically poured into the child's mind merely by subjecting him to the 'right' experiences. But, as we have seen, a person—whether an adult or a child—first conjectures ideas about how the world works, and only then does data and personal experience come to bear as a criticism of those ideas. No amount of sitting in a classroom against their will can guarantee that the young student will receive the intended knowledge from the teacher, and no amount of forcing a child to submit to mandatory rules in the home can guarantee that they learn about the subjects to which the rules pertain. More generally, you can no more force knowledge into a person than you can exceed the speed of light.

Imposed rules aren't just a road to nowhere with respect to children (or any person) acquiring knowledge—they actively *degrade* the child's ability to become a self-actualized, independent person, albeit unintentionally. In particular, the child incurs four costs that he pays for in the form of drained creativity:

1. *The parent-child relationship*: Parents who impose limits on the child's consumption of, say, screens or food necessarily become gatekeepers, enforcers, and judges. If parents do not take measures to prevent the kid

from exceeding these limits, to issue consequences should the kid violate them, and to determine when the kid has violated said limits, then they are not limits at all but rather toothless suggestions. In other words, limits require the parent to act as a kind of homegrown policeman. Far from helping the child learn about the limited thing in question, the child instead learns to regard their parents precisely as gatekeepers, enforcers, and judges, rather than as loving guides that they can trust.

2. *Relationship with self* : Every time a kid has a rule forced on them, it carries with it a negative message about who they are as a person, and this gives the kid a reason to doubt themself. Put differently, there is no way to enforce a rule on a child and guarantee that the child won't take it personally in some way. A rule such as "You mustn't have more than two cookies at a time" signals to the child that his desire to eat a third cookie is somehow wrong, a blemish on his character and preferences. And should he succumb to temptation and eat that third cookie despite the resultant consequences from the parent, he is all the more inadequate.

3. *Confusion about the problem*: Children are too ignorant about the world to live independently, which is why parents have a moral responsibility to steward them until such a time as they are able to continuously solve problems on their own. But rules do not help kids learn about the external world, do not help them foster their personal relationship with such universal and intimate parts of life such as eating, dressing, and socializing. A rule about any of these confuses the child about them, since they are no longer able to freely learn without top-down mediation from the parent. When parents mandate that their children behave a certain way during family meals, discovering the subtleties of dinner table manners becomes discovering how to appease the rule-giver. This does not help the children develop a theory of manners that they can continue to refine via constant conjecture, internal criticism, and feedback from the outside world. Absent such a theory, the child remains ignorant about which aspects of the mandated manners he should modify in novel situations.

4. *Confusion about how to solve problems*: Mandatory rules wrongfully teach children that there are authoritative sources of knowledge. For parenting rules to be effective, it is vital that the children do not consider how the rule might be mistaken, how the parental figure might be wrong. The very paradigm of rules-based parenting, then, wrongly implies that problems

> can be solved by an appeal to authority. But in reality, there are no ultimate, authoritative sources of knowledge or of solutions to the problems we face. As we have said, knowledge creation is an egalitarian enterprise—a child's idea about anything might well contain knowledge that the parent had never before considered.

The truth is that children acquire knowledge and make progress the same way that adults do—by conjecturing solutions to the problems that interest them. Parents certainly have a responsibility to maintain their children's safety and well-being, if only because children lack the knowledge and political freedom to survive and thrive on their own. Neglect is indeed immoral—in fact, the majority of parenting practices and how-to books are themselves neglectful. They neglect children's own reasons and capacity to acquire knowledge via conjecture and criticism. The overwhelming majority of various parenting philosophies, coaches, and how-to guidebooks are ruled out by Popperian philosophy on the grounds that they deny in theory and hinder in practice children's ability to make progress.

The philosophy known as Taking Children Seriously, developed by David Deutsch and Sarah Fitz-Claridge, applies Popper's epistemology to parenting and to the societal treatment of children more generally. Coercive rules cannot work as instruction manuals about the world, and children's lives should instead be full of uninhibited, productive win–win solutions. So, questions such as, "How do I make sure my kid gets into the best college?" and "How do I keep my child grounded in a world full of distractions?" are replaced by, "What are my kid's interests, and how can I foster them in a safe way?" and "Is my child distracted in a way that interferes with his own happiness, or am I presuming to know how he should spend his time?"

For more about Taking Children Seriously, see Conjecture Institute's first book, *The Sovereign Child*, by Cofounder Aaron Stupple with President Logan Chipkin.

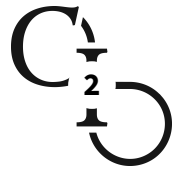
Just as top-down rules cannot facilitate a child's ability to make progress, a standardized curriculum cannot serve to educate large numbers of people—each individual is infinitely unique in his background knowledge, interests, and problem-situations. In light of this Popperian fact, Fellow Chris Sutherland is building an alternative to the bucket theory of mind approach to education that dominates our schools. He is the founder of PhysicsGraph, a mastery-learning platform for the physical sciences, which includes personalized learning paths and built in spaced repetition. His hope is to make education more efficient, effective, and ultimately less coercive.

Popper's epistemology answers many age-old philosophical questions and explains why others are simply misguided. "How are our scientific theories justified?" turns out to be a chimera, since our theories, like all ideas, are simply conjectures. But then, how can we be certain, or probably certain, that our theories are correct? We can't, and, moreover, even our best scientific theories are riddled with errors that future conjectures could resolve. *Certainty* is an impossible aim, but *progress* in science is not.

As for moral philosophy, theories such as consequentialism, deontology, and situational ethics can never be roadmaps, since that would imply that we can discover fixed criteria by which to judge our actions. On the contrary, the very fact that our knowledge is infinitely improvable implies that our moral criteria are as well. And, in any case, our moral ideas are *criticisms* of our actions. We can and should judge our choices according to their consequences, *and* according to whether or not they conform to moral principles that we hold dear. But neither consequentialism nor deontology can possibly serve as a mechanical touchstone that dichotomizes all possible choices between 'right' ones and 'wrong' ones. On the contrary, fallibility implies that our actions grow from 'worse' to 'better' as our knowledge improves from 'more wrong' to 'less wrong'.

Famous thought experiments from moral philosophy, such as the trolley problem, do not reflect actual moral dilemmas that people face in the real world because they do not allow for the creation of new options. In reality, people generate new ideas about how to transform the world around them, thereby expanding their set of choices in the face of moral dilemmas. Morality cannot be distilled to thought experiments in which our choices are fixed. On the contrary, any moral philosophy that does *not* take into account people's ability to fallibly correct errors cannot possibly reflect the human condition.

Domain	Worse Question	Better Question
Politics	Who should rule?	How can we arrange political institutions so that we may peacefully remove and replace rulers and their policies?
Artificial General Intelligence	How can we program an AGI to update its hypotheses' credences in light of new data?	How can we program an AGI to be capable of disobedience and of creating an infinite sequence of ever-improving explanations?
Economics	Which algebraic model best corresponds to an economy?	How can we explain the regularities of an economy in terms of constraints, given that it consists of unpredictable, knowledge-generating people?
Parenting	Which set of rules most guarantee that parents will raise their children to be prepared for today's world?	For any given problem, what are the potential win-win solutions that will satisfy both the parent and the child?
Physics	What is the equation that will unify general relativity and quantum mechanics?	What are the conceptual problems in our current suite of physics theories, and what new explanations might be capable of solving them?
Moral philosophy	Should people be deontologists or consequentialists?	How might we preserve and improve upon our current means of error correction?
Epistemology	How is our knowledge justified?	How does knowledge grow?



Optimism in Action

There's a difference between right and wrong, better and worse, progress and stasis, truth and falsehood. Nearly everyone *says* that these dichotomies are obvious, but few take them seriously.

For instance, if some ideas are more moral than others, and if some choices really are better than others, it follows that the *cultures* that adopt better ideas and choices are more moral than others. Similarly, some cultures really do make progress, while others flounder. In the West, these dichotomies are denied, avoided, ridiculed, and suppressed. Never mind how this came to be—it is a civilizational problem, not only for the West, but for all of humanity.

Western civilization has reached unparalleled heights of prosperity and flourishing. Denying this achievement means never investigating *why*. Specifically, it means failing to discover the institutions, principles, and values that underlie the progress that the West has made. It means taking for granted the necessary ingredients for how a society can transform from a worse one to a better one. It means putting those very ingredients at risk in the West. It means hiding the virtues of the West behind a veil of self-criticism and self-hatred that makes it harder for the non-West to access the same levels of prosperity we enjoy.

Let's deny the taboo-ists and relativists their veto and ask the vital question: why is the West the greatest civilization that the world has ever seen? What are the attributes that distinguish the West from the rest? Once we identify those, we have an understanding of which elements of the West to preserve, which to change, and which to discard. And this understanding not only helps us improve our own society—with the recipe for progress in clearer view, we can, if we so choose, spread them to every other society on the planet. Everybody wins.

The answer must account for two colossal facts: that *errors are inevitable*, and that *mankind was born into ignorance and poverty*. Only in light of this reality does it make

sense that progress of any kind even demands an explanation. (As it happens, many of those who either take the West for granted or deny its superiority do not know these facts.)

If errors are inevitable, then any *particular* error cannot be what distinguishes all other civilizations from the West. The distinguishing factor cannot be, for example, that the West's suite of scientific theories is more advanced than that of the non-Western world. Nor can it be that the West has superior technology as compared with the non-Western world. In fact, non-Western cultures are sometimes *more* advanced in certain domains.

Any innovator will tell you how many mistakes they make before finding success. Counterintuitively, the successful innovator will have made *more* mistakes than the person who chooses a predictable, safe path in life. The same goes for societies: those that make the most progress make the most errors. Ancient Egypt, impressive though its pyramids may be, scarcely improved over thousands of years—the errors of its early years were practically the same errors of its later years. Meanwhile, *because* the West continues to make progress, it necessarily makes entirely novel mistakes on ever-shorter timescales. For example, the United States made mistakes during Reconstruction, but those mistakes would not have even been possible had it not first corrected the error of slavery. Anti-Western critics think that the long list of Western errors is an argument in their favor. On the contrary, it points to the sheer dynamism of the West relative to all of the stagnant cultures that fail to solve problems and thereby make new errors.

The distinguishing characteristic of the West, then, is not a simple accounting of achievements and failures, as non-Western cultures can be more advanced in certain areas, or have avoided this or that failure. What distinguishes the West is that it *corrects errors far better than any historical or contemporary culture*.

The means of error correction, whether within a single mind, the society's culture, a private organization, or a political institution, consist of *processes* that can resolve, improve, or ameliorate an arbitrary sequence of errors.

Each of the West's means of error correction themselves have different attributes, as each evolved to solve a particular problem or collection of problems. The institution of science has peer review, an openness to novel hypotheses, and a university system that, at its best, facilitates both. The institution of political democracy fosters the removal of leaders and policies in favor of an alternative that (some) citizens think will better solve the problems of the day. The institution of cultural norms like freedom of expression, freedom of lifestyle, and freedom of life trajectory facilitate individuals making their own mistakes, learning from them, and changing their minds and choices accordingly. The institution of the market facilitates the allocation and reallocation of scarce resources, as entrepreneurs and consumers alike continuously sell and buy, respectively.

Of course, if errors are inevitable, then we should not expect any of these institutions to be perfect. On the contrary, it is the West and only the West that is currently capable of improving institutions indefinitely.

All of these error correcting institutions have evolved over centuries, some over millennia. And all of them exist primarily as shared ideas across people's minds—if enough people thought that they did not, in fact, correct errors and thus gave up on them, the institutions would disappear in time. On the contrary, if people can explain how and why they really *are* capable of resolving errors, then they will want to retain (and improve upon) them.

In other words, societies that are *optimistic* that errors can be corrected will want to employ and improve upon their means of correcting errors. Societies that are *pessimistic* about the prospects of correcting their errors will allow their means of correcting errors to languish and eventually go extinct.

But choosing between optimism and pessimism is not a matter of taste, mood, or disposition. As David Deutsch argues in *The Beginning of Infinity*, optimism is a physical fact—all errors are, in principle, correctable. It is just a matter of creating the right knowledge of how to do so.

The West is less optimistic today than it has been in the past. This is itself an error that Conjecture Institute exists to correct. Progress in every domain depends on it.

Everything that Conjecture Institute does is a kind of *applied optimism*—we are correcting many of humanity's errors with some of our deepest, yet widely unknown, ideas. Examples include the role of error correction in morality, counterfactuals in physics, freedom in parenting, creativity in artificial general intelligence, knowledge creation in economics, and objectivity in aesthetics.

Ray Scott Percival: Enlightenment of the Mind

Fellow Ray Scott Percival's work applies optimism to thinking itself. The logic and epistemology of the *transmission* of ideas, whether good or bad, is central to his work. In both his documentary, *Liberty Loves Reason*, and his book, *The Myth of the Closed Mind*, he refutes the pessimistic tropes that people are unpersuadable, that minds are indelibly tainted by irrational biases. On the contrary, individuals have been changing their minds since humanity began, and especially since the Enlightenment. And biases are simply conjectures—they may fail, at which point the individual may criticize his own biases and replace them with something better. Or they might foster the discovery of a solution, in which case the biases were not banes to thinking but rather facilitators of it.

Clearing the mind of all biases is not necessarily the right thing to do, since the contents of one's mind must always be understood in light of the problem-situation in which it finds itself. Blind insistence on removing particular errors in one's worldview without intrinsic reason for doing so is a recipe for adopting ideas uncritically. Moreover, not all errors impinge on every problem-situation, so insisting on removing as many errors as you can before trying to solve the problem in front of you is itself an error (and, since errors are inevitable, there is no such thing as an error-free mind).

Typically, your first conjecture will fail, and you will either discard it completely or modify it, and try to solve the problem again. Had your initial conjecture been reliant on some bias, then *maybe* it was this bias that was the reason for the conjecture's failure. But maybe not. Modifying conjectures is always a creative process, and one must be opportunistic when thinking about how to improve them. In short—there is no guarantee that any cognitive error or bias is the reason why a given idea isn't working.

Cognitive biases are not the only errors that can sully a mind. This is yet another reason why clearing the mind of bias is not necessarily the right thing to do—every mind is infinitely unique, with its own collection of errors, knowledge, and interests. Therefore, it cannot be the case that every person's ability to solve his problems is improved by removing cognitive errors. Dogmatically favoring the removal of cognitive errors over all other errors is itself a bias, and one many of us should be aware of—though, according to the very logic of this argument, not everybody.

Ray is now writing another book that will offer prescriptions about how to reignite the Enlightenment in the modern age. He will argue that Enlightenment values, rooted in autonomy, reason, and fallibilism, remain our best hope for human flourishing. Drawing on the ideas of Bartley, Miller, Popper, Deutsch, and Polanyi, the book will critique scientism, defend liberal institutions, and offer a vision of progress grounded in critical rationalism, individual responsibility, and poetic insight.

Arjun Khemani: The Real Story of Humanity

Similarly, Fellow Arjun Khemani's documentary and book, *Lords of the Cosmos* (coauthored by President Logan Chipkin), tell the story of humanity from a knowledge-centered, optimistic worldview. They explain how and why progress took off during the Enlightenment era, which kinds of ideas accelerate and hinder progress, and humanity's role in the cosmic scheme of things. While most 'big history' works focus on events such as the Big Bang, the birth of the first stars, and abiogenesis, the book version of *Lords of the Cosmos* also includes progress-oriented bottlenecks such as the origins of money, the dawn of universal language, and the beginnings of private property. The book also examines *future* knowledge-centered bottlenecks that most futurology works overlook, such as anarcho-capitalism, Taking Children Seriously, and the universal constructor.

Arjun now works to spread awareness of Zcash, a cryptocurrency that is decentralized, private, and scalable. Money is one of the most fundamental technologies of civilization—before the invention of money, trade was conducted through barter. Societies had to match wants directly. Economists call this the problem of the lack of double coincidence of wants. Even though, in principle, anything could be traded for anything, the sheer friction in matching needs made most trades impossible.

Introducing a single good as a medium of exchange (money) solved this. Suddenly, prices could be expressed in a common denominator. And with prices came the possibility of large-scale cooperation. Prices are signals. They convey knowledge about human

preferences across time and space, much as language conveys thought. This is what makes civilization-scale coordination possible.

Throughout history, tyrants have tried to override economic reality by distorting prices. They debased currencies, fixed exchange rates, inflated supply—anything to assert political control over what is, fundamentally, an emergent and decentralized system.

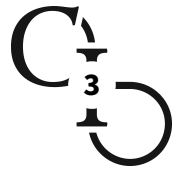
Fiat money is a political artifact. It's fragile, opaque, and prone to abuse. It punishes savers, rewards insiders, and weakens the market's immune system. It makes capitalism unsafe.

Arjun is confident that Zcash may be the solution to all of the problems associated with centralized, inflatable money.

Dimitri Vallein: Cinematic Optimism

Fellow Dimitri Vallein is currently developing a 3D animated short film, *The Day I Met You*, which is yet another example of Conjecture Institute's applied optimism in cinematic form. Whereas so many Hollywood movies tell pessimistic stories of a dystopian future, *The Day I Met You* explores themes of progress through problem-solving, the nature of knowledge creation, and how scientific inquiry can lead to an optimistic future for civilization. Dimitri has already shown his ability to bring scientific and philosophical themes to the cinema with his previous short films, *The Last Star* and *New Specimen*.

Dimitri has also developed a script for a feature-length film, currently titled *Memetics*. The story follows Aria, a young man who creates a simulator to visit ancient societies throughout history. Aria will explore how ideas have shaped and controlled humanity across time and space.



Freedom, Not Outcomes

The unknowability of future scientific theories is a special case of one of the most fundamental limitations in the whole of reality: *the growth of knowledge is unpredictable*. Just as no one could have possibly predicted quantum mechanics by investigating Newtonian mechanics, similarly no one can predict the content of future moral, economic, philosophical, or mathematical theories from our current perch.

In morality, the West has come to regard autonomy, peaceful conflict resolution, dignity, and rights as primary concepts. But a time traveler from the past, certain of his moral standing, would regard such ideas as alien at best and perverse at worst.

In economics, the subjective theory of value emerged only 150 years ago. Had an economist of the tenth century been asked what the future of his field might look like, he very well may have said that humanity would learn of more connections between the Bible and the nature of trade.

Had you asked an ancient Greek what the future of mathematics would entail, he may have contemplated generations of geometers proving an untold number of geometric theorems. "Let no one ignorant of geometry enter" read every potential student upon entering Plato's Academy. But modern mathematicians are spoiled for choices—since the days of Athens' Golden Age, thinkers have conjectured all sorts of fields that no ancient Greek had literally ever dreamed of.

Because of the unity of nature, knowledge in any field is never entirely disconnected from any other. For example, our philosophical outlook has drastically changed radically with many scientific revolutions—sometimes for good, sometimes for bad. Sometimes, bits of our philosophy even go from one position to another and then back to the initial position. As our knowledge of cosmology and biology matured, it seemed that the ancient, religious worldview that placed people at the center of things was laughable. With the advent of constructor theory, we have reversed course—the theory implies that people are, in fact, central to the moral and physical schemes of reality.

Put simply, there is no roadmap, no formula, no criteria that tell us what the content or consequences of future ideas will be. Therefore, questions such as, 'What benefits will your research have for the economy?' are literally nonsensical. *Outcomes in all human endeavors, including scientific research, are unknowable.*

Fixed criteria and mandates imposed on researchers are designed to help funders select between which projects to support. But fixed *anything* is the bane of the very creative process that is required for fundamental breakthroughs, which can come from anywhere, overturn any previously held assumption, and have all sorts of unpredictable downstream consequences on other areas of knowledge. Therefore, researchers who intrinsically *want* to work on risky, bold, potentially revolutionary science will opt for safer, incremental science. With every scientist who makes such a choice for the sake of professional survival, our scientific institutions devolve from a bubbling cauldron of freeroaming ideation into a lifeless, mechanical, 'add another digit to Nature's constants' factory.

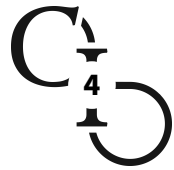
The 'publish or perish' culture similarly distorts the scientific enterprise. Put simply, there is no formula that can tell us the relationship between the number of papers, pages, paragraphs, or words and the number or quality of ideas that a scientist delivers. Insofar as a scientist is pressured to publish for the sake of his career, his papers will tend towards repetitiveness, swathes of content that solve no real problems, and unnecessary graphics.

Turning around the entire academic structure towards the freedom embodied by the Republic of Letters and the Scientific Revolution is a tall order, but modern technology allows us to circumvent it entirely. Thanks to the Internet, our researchers have all the resources they need to develop their academic networks, acquire as much prerequisite knowledge that humanity has to offer as they need, and release their findings first on preprint servers and then in an official journal.

But without the usual academic criteria, how do we decide whom to support? We select researchers who share our particular interests in science, who are realists, who take our best ideas seriously, who are proud to be associated with Conjecture Institute, and who either show promise or who have already solved problems in science that we care about.

But if our researchers are truly free to do whatever they want, how can we guarantee that they'll deliver results? We can't—but then, neither can traditional academia. Imposed criteria don't guarantee outcomes—nothing can. But they do grind the gears of the very progress that the university system purports to care about.

It sounds like a simple model because it *is* a simple model. And it's already working. Several of our researchers have published papers since joining Conjecture Institute.



Realism and Explanations

Progress in science is sluggish. Everyone knows this, but almost no one has the necessary epistemology to explain it. *Conjecture Institute addresses a systemic failure in academia, where groundbreaking ideas often go unfunded due to stultifying norms like incrementalism, empiricism, and explanationless science.*

Scientists regularly present data without any corresponding explanation. Fitting an explanationless equation to data unto itself does not constitute scientific progress. Why *that* equation, and not the infinity of other equations that could just as well have fit the data? Under what other conditions is that equation expected to correspond to physical reality? A paper with such 'results' does not solve a problem in humanity's scientific worldview, but it does solve a problem for the *scientist*—he earns another publication under his belt.

In any case, science is not primarily about equations. Mathematical machinery should be a *tool* of the scientist, not his *goal*. Insisting that an arbitrary bit of explanatory science must correspond to a set of algebraic equations is simply a mistake. Yes, equations have played a vital role in many of our deepest theories in physics and elsewhere, but without a good explanation for why this must continue to be the case, the scientist should be far more opportunistic. He should be open to whatever mathematical objects seem to help him solve his problem—if any at all. Equations work for the scientist, not vice versa.

Relatedly, scientists routinely think that science consists of 'following the evidence', which results in all sorts of funding distortions and wasteful chasing after data. But evidence is not the source of our scientific hypotheses; we cannot 'chase it down' in the hopes of discovering the next great theory. On the contrary, evidence either exposes a problem with our *current* theories, or else it adjudicates between two rival hypotheses (as happens during crucial experiments).

So scientists cannot progress by tacking explanationless equations onto hard-won data nor by 'following the evidence'. Rather, they do so by:

1. Noticing a problem in our suite of best scientific theories (either a conflict between some of the theories, a conflict between data and one or more of the theories, or a contradiction, gap, or vagary within one of the theories), then
2. guessing candidate solutions (which consist primarily of *explanations*), then
3. criticizing all such candidates (including, but not limited to, via crucial experiments), and finally
4. retaining whichever solution has survived all criticisms.

The cultural pressure or internal compulsion to prioritize data, graphs, and equations causes scientists to skip (1) - (3) and proclaim that they have discovered a solution to a problem. But far too often, the problem wasn't a real scientific problem, and their 'solution' was not an explanation of anything.

(1) Requires that scientists take our best scientific theories seriously as explanations about reality, rather than as merely calculation machines. Newton would never have been bothered by the conflict between the principle of locality and his theory of gravity if he had viewed his work as just a set of mathematical instruments that could make impressive predictions. Count Rumford would never have conducted experiments to determine whether heat was a fluid or not if he had only cared about calculating the efficiency of thermodynamic engines. Einstein may have foregone wondering what it might mean to take the principle of relativity and the constancy of the speed of light seriously and instead searched for a mathematical object that might contain the apparently contradictory Newtonian mechanics and Maxwell's theory of electromagnetism.

The current trend of treating theories as merely calculation machines is intimately connected to the other bad trend of looking for solutions to *physical* problems in *mathematics*. It is a kind of 'mathematics in, mathematics' out philosophy. But physical problems require physical solutions. Newton's theory worked just fine, yet he was dissatisfied with the apparent nonlocal effects that gravity had *in physical reality*.

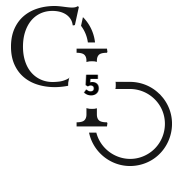
How our current scientific culture came to treat theories as anything less than explanations of the physical world is a project for a historian. In the meantime, Conjecture Institute is doing something about it—we support precisely those physicists who accept that our theories really do say something about the world. Infrared light really does exist, even though we cannot see it. Quantum information really does exist, even though it is an abstraction. The quantum multiverse really does exist, even though the world of everyday experience suggests otherwise. And, most importantly, the world really does make sense, and problems in our scientific worldview are there to be solved.

(2) is hampered by a culture that, implicitly or otherwise, insists that the next fundamental theory must have some features that previous fundamental theories had. Some scientists insist that the next revolutionary theory in physics will be in terms of laws of motion and initial conditions. Others insist that it will be in terms of ever

smaller objects. Still others insist that it will be characterized by a Lagrangian. All research programmes that conform to these criteria assume, either by fiat or unstated assumption, that the next theory *must* possess one or more of these comfortably familiar characteristics.

But it is impossible to know in advance which aspects of the previous theory will survive the next scientific revolution fully intact and which will be shown to be merely limiting cases of a deeper regularity that is expressed in an entirely new conceptual language. Had the inventors of quantum mechanics insisted that it be expressible in terms of Newtonian concepts like single-valued quantities, they'd have never arrived at their ideas. On the contrary, modern quantum mechanics relies on counterintuitive ideas like interference, entanglement, and the multiverse. Mathematically, quantum mechanics has in common with Newtonian mechanics that differential equations are central, yet those of quantum mechanics include matrices, while those of Newtonian mechanics do not.

Constructor theory, a new research programme in fundamental physics that Conjecture Institute supports, is a prime example of a theory that does not insist on taking the form of the fundamental theories that have come before it. And yet, amazingly, Conjecture Institute Advisor David Deutsch, Conjecture Institute Senior Scientist Chiara Marletto, Conjecture Institute Senior Scientist Vlatko Vedral, Conjecture Institute Fellow Maria Violaris, Conjecture Institute Fellow Giuseppe Di Pietra, and others have already used it to solve real problems in physics.



Further Applications

Tom Hyde: The Fusion of Reason and Beauty

If beauty is objective, then we really can make progress in the arts, both theoretically and practically. Theoretically, this means that our criteria for what we think is beautiful are improvable, and that our theories of beauty can be made more sound, more mature, and more constraining. Practically, this means both that our artistic creations are improvable within a single medium (i.e., some movies and songs are objectively better than others), but also that there exists an infinite sequence of possible *modes* of artistic creation, some better than others. As we create new technologies, and as our moral and scientific knowledge improve, we correspondingly acquire the capacity to tell better stories and simulate better experiences.

Beauty is as real as science, reason as real as romance, but what is the relationship between them, if any? There is surely *some* connection—it is not an accident that physicists regularly describe Einstein's theory of general relativity as beautiful, nor that authors and painters alike take inspiration from our deepest scientific ideas. If there exists some unifying theory that constrains the relationships between art and science, reason and wonder, then we can exploit them as much as computer engineers exploit quantum mechanics and photographers exploit compositional theories. In other words, a unifying theory of science and the arts would accelerate progress in both.

Conjecture Institute Fellow Tom Hyde is currently developing his theory of Wonderism precisely to address this understudied domain. Specifically, he aims to unify the emphasis on reason, thought, logic, and sense that characterizes the scientific subculture of the Enlightenment with the emphasis on passion, feeling, instinct, and sensibility that characterizes Romanticism. Whereas the Enlightenment values utility and function, Romanticism values beauty and form.

Tom conjectures that there is not a conflict between Enlightenment's Age of Reason and the artistic movement of Romanticism. Traditionally, the Enlightenment tells us not to feel but to think, and Romanticism says not to think but to feel. Wonderism, meanwhile, tells us that in order to think, we must feel, and that in order to feel, we must think.

Wonderism adds further unifying constraints on Enlightenment and Romantic thought. For example, Wonderism posits that the *curiosity* felt by scientists in pursuit of a discovery and the *adoration* felt by consumers of art are, in fact, both necessary ingredients in science and art. Scientists necessarily adore a good solution to their problem, and artists are necessarily curious as they explore the space of their next possible artistic stroke (or sentence, or storyline, or musical key).

Wonderism is the confluence of the two intellectual movements, where the seemingly distinct concepts of *wonder* from reason and passion become one. The scientist thoughtfully wonders what matter is made of, the size of the universe, how a star is born. The artist channels inspirational wonder upon peering off the edge of a mountaintop, smelling the damp earth of a forest, dipping her toe in the roaring vastness of the ocean. Wonderism tells us that these two senses of wonder are, in fact, the same, and that the same processes cause and are caused by both.

In Wonderism, the two wonders build on one another: satisfying Enlightenment wonders—those grounded in science and reason—opens the door to exploring never-before-conceived Romanticism wonders—those grounded in experience and art. Those Romantic wonders, in turn, will illuminate and beautify new problems in our worldview that Enlightenment wonderers will then endeavor to solve. If there are any more specific constraints to this synergistic cycle, then they will be expressed and discovered in terms of Tom's Wonderism.

Carlos De la Guardia: Artificial General Intelligence without the Pitfalls

Conjecture Institute Fellow Carlos De la Guardia 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?

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 -Reachable.

Active ideas are those ideas that we are currently applying to whatever problem we face. Existing ideas are a much larger superset of active ideas, consisting of all of the knowledge that constitutes a given mind. N-Reachable ideas are those not currently

instantiated in a mind but could become so in N steps. -Reachable ideas are all of the ideas that a mind could ever reach, given unlimited time and resources. Carlos conjectures that the set of all -Reachable ideas for any individual mind corresponds to the set of all possible ideas that any other mind could acquire.

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 -Reachable ideas.

Another insight that has come out of Carlos' research is the five ways by 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's little good having lots of ideas if you can't apply the right ones in the right problem situations. If you stub your toe, it's 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.

Carlos' approach to artificial general intelligence is expressible in purely epistemological and physical terms. To be sure, at some point, this will have to be translated into a software program. But, 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.

Paul Raymond-Robichaud: Proving Constraints on Future Theories in Physics

As we have seen, to instrumentalize a scientific theory is to not take it seriously as an account of what the world is really like. Another common way by which scientists do not take our best ideas seriously is by denying their universal character. For example, physical principles are known to constrain all theories in physics: the principle of conservation of energy, the Church-Turing-Deutsch principle, and the principle of locality apply to all theories in physics, both known and unknown. Yet when there is an apparent contradiction between, say, the principle of locality and quantum mechanics, many physicists suggest that the principle of locality might allow for an exception. Taking the principle of locality seriously, on the other hand, would mean figuring out a way to express quantum mechanics in a way that is fully consistent with the principle.

Conjecture Institute Fellow Paul Raymond-Robichaud has proven that *any* physical theory that supports a no-signaling theorem and has reversible dynamics must have a local-realistic formulation. In layman's terms, Paul has demonstrated that if a theory in physics forbids the operation on one system having an instantaneous observable effect on remote systems *and* is characterized by equations of motion that can be reversed, then that theory necessarily conforms to the principle of locality. So it's not just quantum theory that is, in fact, local—even as-yet discovered theories, provided that they conform to the aforementioned criteria, will also be local.

Paul's work has rendered obsolete so much current research in quantum theory: scientists insist on *non*-locality when investigating entanglement and Bell's theorem,

thereby wasting precious resources going down blind alleys. It would be as if chemists insisted that the law of definite proportions is not universal for all chemical reactions and insisted instead that, in some cases, we might explain a chemical reaction via other means. But because the law of definite proportions is, in fact, universal for all chemical reactions, the chemists will unnecessarily run into a brick wall for the singular reason that they deny taking our deepest constraints seriously.

Relatedly, Paul's work sheds light on the decades-long battle over the correct *interpretation o*f quantum mechanics. In particular, Paul's proof rules out local-hidden variable formulations of the theory as, for example, introduced by John Stewart Bell. Many scientists think that the reality of entanglement, a situation in which the state of a system cannot be deduced from the state of its subsystems, implies that quantum theory violates locality. But because of Paul's work, we now know that entanglement is perfectly compatible with locality, and that making this explicit is 'merely' a matter of discovering the appropriate formalism.

Moreover, because Paul's work applies not just to quantum theory but to an arbitrary sequence of as-yet unknown theories, his proof is vital in the pursuit of, say, a theory of quantum gravity. Just as we expect the true theory of quantum gravity to conform to the principle of conservation of energy and, therefore, we may immediately rule out any hypothesis that contradicts it, so too with Paul's theorem. That is, if a conjectured theory of quantum gravity contradicts the principle of locality, then, provided that it also supports a no-signalling theorem and has reversible dynamics, it cannot possibly be a valid theory.

Fellow Samuel Hagh Shenas' course for Conjecture University, *The Principle of Locality*, explores the history of the principle of locality and the role it has played in our understanding of the physical world. Because locality transcends any particular theory in physics, it comes up again and again in different contexts: Newton, Maxwell, quantum physicists, and Einstein all dealt with it in subtly distinct and illuminating ways.

The Unreality of Time

Time is similar to locality in that its existence transcends any particular theory in physics and constrains all of them, both known and unknown. To contrast, the statistical behavior of gas particles has no bearing on the tenets or implications of general relativity, and the no-cloning theorem of quantum mechanics has no bearing on the laws of thermodynamics. And yet time plays an (apparently) inextricable role in general relativity, quantum mechanics, *and* thermodynamics (in general, time plays a role in any theory that describes change). The physics of time, then, has unavoidably wide implications for our understanding of the physical world.

Yet time has an awkward place in physics, at least as fundamental theories are typically expressed. It is invoked as a real-valued parameter relative to which systems change, yet the time parameter itself is not physical. That is, while *other* attributes of physical systems exert physical effects on the world and are exerted upon in turn, the time parameter is

treated as an unphysical abstraction that has no role other than to give chronology to the sequence of states of a system and its attributes. One hoped-for solution would be to excise this ad hoc parameter from physics entirely, leaving time as a merely emergent, approximative feature of physics that can be done away with, and not one that is intrinsic to the evolution of systems.

Physicist Julian Barbour has demonstrated that the dynamics of Newtonian systems can be modeled without invoking an external time parameter. In his model, a collection of point masses' configuration in space can be uniquely characterized at any instant, thereby rendering any reference to time obsolete and redundant. In other words, Barbour showed that it is possible to quantify the evolution of a Newtonian system by labeling and ordering the various states of the system without ever needing to make calculations in which time is an input parameter.

Can time be similarly rendered obsolete for quantum mechanical systems? Time is notoriously problematic in quantum theory, since every observable besides time corresponds to a matrix, while time alone is treated as a real-valued number—a glaring blemish on an otherwise elegant mathematical framework. Furthermore, this leaves a gap between theory and experiment, since quantum systems themselves are routinely used as clocks to measure time in an experiment, and yet it is not known how to treat time the same way that all other measurable observables in quantum mechanics are treated.

The Page-Wootters construction had partially solved this discrepancy, as it modeled the universe as being in a stationary state that evolves relative to a 'clock' inside of it. While this eliminated the need for an external time parameter, the assumption that the clock is isolated prevents it from being measured, which is the whole point of excising external time from physics in the first place. Conjecture Institute Fellow Sam Kuypers (in collaboration with Simone Rijavec) proved that, indeed, even a non-isolated clock can be measured within the Page-Wootters construction without invoking traditional conceptions of time. As Barbour had done for Newtonian gravity, Sam and Simone have demonstrated that the evolution of quantum systems can be characterized and measured without invoking any awkward, unphysical time parameter.

If every physical theory has a formulation in which the time parameter is absent, then one might expect that the aforementioned constructor theory, itself a theory that purports to underlie all other theories, might shed light on a timeless formulation of the entirety of the laws of physics. Indeed, the central tenet of constructor theory, that the laws of physics are expressible in terms of possible and impossible tasks, already implies that physical laws must be expressible in timeless terms. A task is either possible or impossible, not 'possible in a given duration t' or 'impossible with duration less than time t'. Therefore, if a law of physics is expressed in constructor theoretic terms—the mathematical formalism of which is called task algebra—then that law will automatically and simultaneously be expressed in timeless terms, i.e. in terms that do not include an extrinsic time parameter.

Yet if our deepest theory in physics allows for a timeless formulation of the laws of physics, how and why do duration and dynamics (the evolution of systems over time) emerge in the first place?

In a recent paper, Conjecture Institute Advisor David Deutsch and Senior Scientist Chiara Marletto addressed this problem, showing how one can recover dynamics as emerging from timeless principles of constructor theory. They also explain the regularities in nature that are necessary for timeless approaches to be possible.

The Unreality of Probability

The central tenet of constructor theory immediately implies that the laws of physics are deterministic, not probabilistic. A task is either possible or impossible, not 'probably possible' or even 'possible with probability '. Like time, classicality, and the flat Earth theory, probability is an approximation that can be useful in some situations but a bane against progress in science if taken too seriously.

But isn't quantum mechanics fundamentally probabilistic? No—the unpredictability and appearance of stochasticity of measurements are completely compatible with an entirely deterministic version of quantum mechanics.

Marletto has applied the constructor theory of information (see below) to the problem of explaining the conditions under which the unpredictability and appearance of stochasticity of measurements can emerge in a deterministic world. Her argument generalizes beyond quantum theory and applies to any theory that allows for so-called superinformation and therefore is expected to apply to whatever theory supersedes quantum mechanics.

Deutsch has explained how one can excise probability from disparate fields such as evolution, the theory of experimental errors, quantum theory, information theory, and Bayesian philosophy of science.

The Heisenberg Picture of Quantum Mechanics

In addition to taking locality and timelessness seriously, yet another avenue for progress in physics is that of the Heisenberg picture of quantum mechanics. Traditionally, physicists express the equations of quantum mechanics in terms of the Schrodinger picture, in which the state vector evolves and the observables are fixed. In the Heisenberg picture, the state vector is fixed while the observables evolve over time. Naively, one may think that the two mathematical formalisms offer physically equivalent descriptions of reality, since they do, in fact, yield the same predictions. However, as Fellow Charles Bedard has explained, this equivalence is only satisfactory if one takes an instrumentalist view of our theories, rather than a realist one. If one takes our theories not as merely prediction machines but as actual descriptions of the world, then the Schrodinger and Heisenberg pictures are not at all equivalent.

Bedard further explains why the Heisenberg picture offers a richer ontology than the Schrodinger picture. For example, it allows an explicitly local account of superdense coding, teleportation, branching, and Bell inequality violations—phenomena that the Schrodinger framework does not explain in fully local terms. Furthermore, unlike the Schrodinger picture, the Heisenberg picture allows for transparent analysis of information flow in quantum systems.

Charles Bedard: Frontiers and Foundations

Charles is also in the process of creating a Montreal-based, Conjecture Institute-affiliated laboratory for physics, informatics, and philosophy tentatively called 'Frontiers and Foundations'. The laboratory will publish, lecture, and collaborate across academic and industrial boundaries, deliberately ignoring the lines that separate departments and traditions. It views science as a living conversation about reality itself, not as a collection of techniques.

Frontiers and Foundations will abide by principles that are completely aligned with Conjecture Institute's broader ethos:

1. *Ignore artificial boundaries between disciplines*: the deeper our ideas go, the more integrated our knowledge of reality becomes.
2. *A culture of criticism*: because Charles is deeply influenced by critical rationalism, his lab will adopt a culture of good-natured criticism that no idea may rise above.
3. *Curiosity over 'utility'*: researchers in Charles' lab will pursue ideas in which they are intrinsically interested, not whichever ideas that they think will have the 'greatest impact' on society.
4. *Open discussions*: seminars, workshops, and media appearances will be publicly available whenever possible. Furthermore, every result will be published in paper, blog, or video form for rapid public feedback.

All research at the Frontiers and Foundations lab revolves around a single ambition—to explain how causation, classicality, and understanding itself emerge from the quantum world:

1. *Causality in a quantum world*: researchers will examine how causal structure, temporal order, and non-influence conditions emerge from unitary dynamics (the time evolution of a quantum system), as well as how to bridge descriptor-based causality (causality in the Heisenberg picture) and modern causal models (such as those investigated by Conjecture Institute Advisor Judea Pearl).
2. *Heisenberg picture*: researchers will investigate how quasi-classical regimes arise without collapse or observers, how relative descriptors encode stable classical information, and how the Heisenberg state participates in classicality, decoherence, and the emergence of robust facts.

3. *Epistemology as a physical theory*: drawing on algorithmic information theory, universal computation, and connections between universal constructors (theoretical machines that can be programmed to cause any physical transformation) and universal explainers (entities—namely, people—that are capable of explaining everything that can be explained), they will explore how explanation, understanding, and model-building can be given physical expression. They aim to establish epistemology as a natural extension of physics, rather than an after-the-fact commentary on it.

Taking Quantum Theory Seriously: Applications and a New Textbook

The aforementioned *branching* is another concept that is controversial in the academy, but only because of the pervasiveness of instrumentalism and empiricism. The only interpretation of quantum mechanics consistent with realism is the Everettian, or multiverse, interpretation. Rather than add an ad hoc rule to the theory that the wavefunction collapses upon measurement, one instead takes the 'bare' rules of the theory and infers that multiple occurrences of a given measurement take place across the quantum multiverse.

Conjecture Institute Fellow Maxime Desalle explains why the Everettian interpretation makes sense in his course for Conjecture University, titled *Taking Schrödinger Seriously*. Designed for a lay audience but without watering down the ideas, Maxime's course investigates the central equation in quantum mechanics and explains the implications for taking it seriously as a reflection of what reality is actually like.

Sam Kuypers has moved physics forward by casting quantum physics in terms of both the Heisenberg picture and the Everett interpretation—sometimes simultaneously. For example, Kuypers and Deutsch broke new ground by expressing Everett's relative-state construction in the Heisenberg picture, as it had previously only been expressed in the Schrodinger picture. The result was a construction that, unlike Everett's one in the Schrödinger picture, makes manifest the locality of Everettian multiplicity, its inherently approximative nature, and its origin in certain kinds of entanglement and locally inaccessible information.

As we have seen, the Heisenberg picture and the Everett interpretation are both drastically underappreciated in the halls of academia, yet they have been and will continue to be vital for making progress in fundamental physics. The overwhelming majority of quantum mechanics textbooks downplay the Heisenberg picture and equivocate on the Everett interpretation at best and endorse an instrumentalist view of quantum physics at worst. They also tend to use unnecessarily complicated examples such as, say, the hydrogen atom and harmonic oscillators that require the reader to grok concepts and mathematical superstructures that are not intrinsic to the principles of quantum theory.

Imagine if our best textbook on evolutionary theory today was written before the revolution of the selfish gene model, or if our best textbook on special relativity was written before the Minkowski spacetime was invented. That is precisely where we are with respect to quantum mechanics textbooks—there is scope for an update.

To that end, Deutsch, Marletto, and Kuypers—all Conjecture Institute affiliates—are writing a quantum mechanics textbook that fully incorporates the latest developments on the Heisenberg picture, the Everett interpretation, and quantum information theory.

More Progress in—and with—Constructor Theory

Constructor Theory of Information

Until constructor theory, the regularities of information had not been integrated into fundamental physics. For one thing, the physics of information is not bound to any particular theory, such as Newtonian mechanics, thermodynamics, or quantum computation. On the contrary, some theories seem to constrain the kind of informational transformations that are possible (such as quantum computation), while others consist of principles that seem to refer to information directly (such as thermodynamics). And yet an autonomous theory that captures the regularities of information, independent of references to other physical domains, has eluded scientists.

The regularities that characterize information are scale-independent (qubits and bits can be embedded in objects from photons to atoms to molecules), counterfactual in nature (a red sign that is capable of switching to blue or green conveys information, even if it happens to never change color), and substrate-independent (the same song can exist in, and be transferred between, sound waves, a human brain, and an electronic recording device). All three of these characteristics are difficult or impossible to capture in the prevailing conception but are entirely natural in constructor theory. Therefore, it is perhaps not surprising that one of the first solutions that constructor theory delivered was the constructor theory of information—the first and only robust theory of information that integrates it into the rest of physics.

When investigating the physics of information, it is important to remember that information is not *prior* to physics—like mass, energy, motion, and structure, it is the laws of physics that determine, explain, and constrain its regularities, not vice versa. Many attempts to explain information had mistakenly assumed that information was, like the regularities of mathematical objects, prior to the laws that govern our universe.

Shannon's theory of information, what had been the prevailing theory of information, was problematic in at least two ways. Firstly, it fails for quantum information—it does not, for instance, take into account prohibitions like no-cloning and noncommutativity that characterize quantum information. Secondly, it does not specify what distinguishability—a critical concept in any information theory, as it is necessary for

characterizing the concepts of measurement and preparation in turn—consists of in physical terms.

Quantum theory of information, though it has been fruitful in its own right, is not actually a fully fledged theory but rather a collection of quantum phenomena that violates the laws of classical information. Consequently, it does not explain its relationship to classical physics, nor can it possibly serve as a theory of information for whatever theory supersedes quantum mechanics.

Constructor theory of information solves all of these problems wholesale, providing a physical characterization of distinguishability, explaining the key difference between classical and quantum information, capturing the regularities of information in terms of scale-independent, nonprobabilistic principles.

Now, what *is* the difference between classical and quantum information? Amazingly, all of the phenomena that we associate uniquely with quantum information come out of a single constructor theoretic idea: a superinformation medium is an information medium with at least two information observables that contain only mutually disjoint attributes and whose union is not an information observable. From this singular, additional prohibition, the following familiar properties of quantum information come out: the undetectability of sharpness, the impossibility of cloning, the impossibility of simultaneous preparation or measurement of pairs of observables, the unpredictability of deterministic processes, and others.

To be sure, quantum information is but an instance of superinformation, and the constructor theoretic principles of superinformation will constrain and apply to all theories that supersede quantum information, just as the principle of conservation of energy and the principle of locality will.

An Unambiguous and Cost-Effective Test of Quantum Gravity

One of the first applications of the constructor theory of information has been in the development of the BMV (Bose-Marletto-Vedral) experiment.

One of the greatest outstanding problems in science is the clash between quantum theory and gravity: each is enormously successful in its respective domain, yet they are inconsistent with each other in several ways. For example, Einstein's theory of gravity describes the world as *classical and deterministic*, while quantum theory tells us that indeterminacy of measurement, superpositions, and entanglement are fundamental aspects of reality.

A number of proposals have been submitted to unify these two worldviews, such as loop quantum gravity or string theory. But all such proposals are practically impossible to test.

By applying the constructor theory of information, Conjecture Institute Senior Scientists Vlatko Vedral and Chiara Marletto have already discovered some of the characteristics that a unified theory of quantum gravity must possess. They have demonstrated that *if* gravity can mediate entanglement between two quantum systems, *then* gravity

itself must be quantum in the sense that it is characterized by at least two observables that cannot be simultaneously measured to arbitrary accuracy and can transmit quantum information. Their argument is independent of the particular dynamics of either quantum theory or general relativity—they assume only the principle of locality and the constructor theory of information.

Testing the claim that gravity is quantum via the BMV experiment is far more feasible than testing any popular theories of quantum gravity. While erecting and running the LHC has cost billions of dollars, the BMV experiment costs about $7 million. It would also be far more fruitful than yet another desperate search for new particles.

If the BMV experiment succeeds in demonstrating that gravity can mediate entanglement between two quantum systems and is therefore itself quantum, then all proposed theories of quantum gravity that insist on gravity remaining classical are immediately ruled out. For example, all so-called semiclassical theories of gravity are refuted. In these theories, the background spacetime remains classical.

The BMV experiment would also rule out collapse-type models that predict a collapse of the wavefunction of each mass at the scales of the experiment.

Finally, the BMV experiment would rule out proposals that treat gravity as an induced field by the quantum vacuum fluctuations of all other fields. According to this logic, gravity is not a fundamental force and therefore need not be quantized at all.

Conjecture Institute Fellow Antonia Weber, in collaboration with Vlatko Vedral, has done theoretical work on aspects of the BMV experiment. For example, she has shown that, in fact, *all* of the components of the gravitational potential must be non-classical should the experiment demonstrate entanglement.

Following the execution of the BMV experiment, and assuming that it indeed demonstrates gravity's quantum nature, theorists can then begin to investigate in full what it means for gravity to be quantum: Which of its components behave like quantum observables? Which cannot be measured simultaneously? Which can transmit quantum information?

Witnessing Non-Classicality in Biological Systems

Another application of the constructor theory of information is in biology. Fellow Giuseppe Di Pietra has developed a strategy to test whether a particular biological process (photosynthetic energy transport) is classical or quantum, which has been notoriously difficult to determine in the history of quantum biology. The strategy is to introduce a second system: a *quantum probe* whose behaviour one can control and measure with high precision to test the non-classical properties of the investigated process. The probe interacts with the biological system in such a way that any non-classical features of the biological process are imprinted onto the probe itself. In this sense, the quantum probe acts as a detective, investigating whether the process is quantum or not without disturbing it. If the host biological system is genuinely quantum, it can mediate non-classical effects on the quantum probe that would otherwise be unattainable. If the

probe exhibits such effects *under appropriate conditions*, then *all* classical descriptions of the mediating process must be ruled out.

Constructor Theory of Life

A different application of the constructor theory of information was to a longstanding problem in theoretical biology: the design of biological adaptations is not encoded in the laws of physics, and yet Neo-Darwinian evolutionary theory relies on the possibility of certain physical processes, mainly gene replication and natural selection. What properties must 'no-design' laws of physics have to allow for such processes?

Chiara has applied constructor theory to provide an exact formulation of the appearance of design, of no-design laws, and of the logic of self-reproduction and natural selection within fundamental physics. She has further demonstrated that self-reproduction, replication, and natural selection are possible under no-design laws, the only non-trivial condition being that they allow digital information to be physically instantiated. This condition has an exact characterization in the constructor theory of information.

Constructor Theory of Thermodynamics

Problems in thermodynamics were a natural early target for the constructor theory research program, as traditional formulations were already—either implicitly or explicitly—expressed in terms of possible and impossible transformations. Moreover, one issue with traditional thermodynamics is that it is scale-dependent—the laws hold at a certain scale or level of coarse-graining that is never specified. In practice, this means that, while the laws of thermodynamics can and have been applied to macroscopic objects such as Victorian heat engines, they could never be applied to microscopic systems such as nanoengines or individual quantum systems. If the future of technology is microscopic, then we will need to know how to, for example, distinguish between work and heat at arbitrarily fine scales.

The traditional formulation of the zeroth law of thermodynamics posits that if two systems are in thermal equilibrium with a third system, then they are in equilibrium with each other. However, equilibrium conditions are themselves scale-dependent. For example, if equilibrium states are defined as those "which, once attained, remain constant in time thereafter until the external conditions are unchanged", then they are *never* attained, if only due to classical and quantum fluctuations. In other words, the equilibrium as understood in traditional thermodynamics is merely a sometimes-useful fiction. And because the first and second laws of thermodynamics also invoke equilibrium, this problematic logic extends to them as well.

In Chiara's constructor theoretic formulation of thermodynamics, the zeroth law is not fundamentally about equilibrium or temperature but rather about the possibility of certain tasks that can or cannot be rendered at *any* scale. It states that, given any thermodynamic attribute x, it is possible to transform x into any heat attribute h with

the same energy as x. This formulation applies to single-particle systems, systems in or out of equilibrium, and systems whose states can be brought about either spontaneously or nonspontaneously.

The first law of thermodynamics states that all ways of performing work are equivalent in the sense that the same amount of energy is required to perform a given amount of work, regardless of the form that the energy takes (i.e. electrochemical, gravitational, electromagnetic, nuclear, etc.). The constructor theoretic version of this law is both more exact and more general, stating that work media are interoperable with one another. This formulation, paired with the constructor theory of information, has revealed a novel connection between the physics of information and that of thermodynamics: *if* it is possible to extract work deterministically from each of a set of attributes of a physical system, *then* the attributes in that set are all distinguishable from one another. This scale-free theorem applies to classical systems, quantum systems, and even hybrid systems (i.e. those at the boundary of quantum mechanics and general relativity).

The statistical mechanics approach to the second law of thermodynamics casts it as requiring there to be irreversibility in some spontaneous evolution of confined, isolated systems such that entropy increases globally—informally, that disorder always increases in the long run. But this is notoriously incongruent with the fact that the dynamical laws of motion are time-reversal symmetric. For example, the Newtonian laws of motion allow for both a plate to fall to the floor and crack into dozens of pieces and the reverse happening. And yet, consistent with common sense, we only ever observe the former and never the latter. Some models of physical systems have irreversibility emerging from underlying symmetrical dynamics, but they always rely on one approximation or another—either applying only under select (and ad hoc) conditions, appealing to some trajectories being more probable than others, or only working given certain initial conditions.

Chiara's constructor theoretic formulation of the second law is not primarily about entropy or spontaneous evolution. Instead, Chiara's version of the second law is about the fact that some tasks are possible in one direction but impossible in the reverse direction. Unlike previous formulations of the second law, the constructor theoretic one is perfectly compatible with time-reversal symmetric dynamical laws. This 'law of impotence' version also renders the emergence of the arrow of time and the second law of thermodynamics as entirely distinct issues, since, as we have seen, constructor theoretic laws are always expressed without reference to time.

Maria Violaris: Communicating Physics with Thought Experiments

Thought experiments have played a crucial role in the history of science, from Maxwell's demon to Schrodinger's cat to Einstein's light beam. Thought experiments only make sense when you take ideas seriously—after all, if you mistakenly view a theory as merely

a prediction machine, how could you ever contemplate what might happen under these or those conditions? Maxwell thought that there really was such a thing as information. Schrodinger saw what the world looked like if quantum mechanics was taken seriously via his famous cat being alive and dead simultaneously (alive in some worlds and dead in others). Einstein thought that light really did move at constant velocity for all observers, even those running near the speed of light.

Conjecture Institute Fellow Maria Violaris conducts original physics research and communicates established ideas (mostly from quantum information theory) to a lay audience. Inspired by the power of thought experiments throughout the history of science, she has written an educational paper about many of quantum physics' most famous thought experiments, from Schrodinger's cat to Deutsch's futuristic test of the Many Worlds interpretation to Wigner's friend (this highlights the measurement problem). Maria has created quantum circuits that correspond to each of these thought experiments, both to demystify the seemingly paradoxical nature of some of them, and to provide readers with further examples and tools to learn quantum computation. This paper was also inspired by Maria's 'quantum paradoxes' content series that she had previously created across video, blogs, and code tutorials.

Maria later wrote a follow-up paper with more thought experiments meant to advance the reader's understanding of quantum computation, special relativity, general relativity, and thermodynamics. Thought experiments in this sequel paper include: the EPR experiment, Maxwell's demon, and the grandfather 'paradox'.

Building on her previous work, Maria has drafted an outline for a new series of videos and/or book that explores nine thought experiments that have informed our understanding of science, philosophy, and technology:

1. Iconic quantum thought experiments (Maria will explain why the notorious 'spooky action at a distance' is not so puzzling, and why apparent paradoxes are perfectly self-consistent):
2. Schrodinger's cat
3. Double slit experiment
4. Einstein-Podolsky-Rosen (EPR) paradox
5. Thought experiments that fuse conceptual breakthroughs with entirely new technologies:
6. Bomb tester
7. Many Worlds vs. collapse
8. Quantum teleportation
9. The world of quantum spacetime (by exposing the contradictions that arise in each of these thought experiments, Maria will highlight crucial clues that they give us towards finding future theories):

10. Time-loops
11. Quantum gravity
12. Black hole information paradox

Eric Denton: Explaining Epistemology to the World

As you have seen, the majority of our Fellows work at the frontier of human knowledge. However, it is also important to review and explain the ideas that have brought us here. Fellow Eric Denton is a writer and podcaster who is passionate about spreading the core ideas of critical rationalism, which he does through his podcast, *The Falsifiable Podcast*. His upcoming book will be an exploration of the lives and ideas of Karl Popper and Charles Darwin, and the significance of how and why their theories overlap.

Brett Hall: Communication and Extending Our Deepest Ideas

Conjecture Institute Ambassador Brett Hall similarly explains this worldview through his writings, podcast (TokCast), and public debates and discussions with prominent intellectuals. Brett also elaborates on the ideas and applies them in original and insightful ways. For example, in his recent book (and Conjecture Institute's fourth), *The Farthest Reaches*, Brett takes the single idea of explanatory universality as introduced in Deutsch's *The Beginning of Infinity* and applies it to several current issues and controversies.

Explanatory universality is the idea that there exists a single object that can, in principle, explain anything that can be explained. *People*, of which humans are an instance, are precisely such objects. Children, artificial general intelligence, and advanced aliens also possess explanatory universality—they are all *people*.

In *The Farthest Reaches*, Brett takes this idea and applies it to, among other things:

1. Education
2. Psychological science and IQ
3. Transgenderism
4. Immigration
5. Ethnicity
6. Multiculturalism
7. Mental illness
8. Sexuality
9. Evolutionary psychology

Doing History Right

The discipline of history stands to benefit from Deutschian critical rationalism in a number of ways. First of all, the history of humanity is not the history of material conditions, nor is it a Manichean struggle between oppressed and oppressor or colonized and colonizer or bourgeoisie and proletariat, nor is it about only select heroic and villainous individuals who supposedly determine the course of humanity. History is about *ideas*: how the material conditions affect any particular society is determined by their culture, institutions, worldview, and memes. Purportedly Manichean struggles between groups can never be permanent features of humanity—any given struggle is due to conflicting views about how people should interact and coordinate, and these views can be improved upon and converge across all parties. And the strongman view of history discounts the role of the gradual, bottom-up evolution of ideas that drives the nameless masses.

If history is driven primarily by ideas, then the historian has no choice but to take the institutions of the historical period of interest into account. Institutions, after all, are catalysts composed of knowledge whose purpose is to foster the propagation of ideas that wouldn't be possible in their absence. The historian, then, must incorporate all of the institutions' relevant attributes into his explanation of events—not just the explicit ones, nor the most salient ones, nor the ones most appreciated during said historical period.

For example, there is a lot more to a religion than its holy texts. First of all, every adherent will *interpret* the texts idiosyncratically, based on the background knowledge they bring to bear. Long-lasting religious schisms have erupted over precisely this phenomenon. Secondly, even if adherents do converge on an interpretation of the text's mandates, they may diverge on how to *adhere* to them—that is, on how to integrate their doctrine into their lives. Analysis of the Bible alone is insufficient to explain Christianity's role in the history of humanity, if only because one must also explain how the Bible was interpreted and implemented by Christians at any point in history. In an open, dynamic society, a historian should expect that the background knowledge that adherents bring to their holy texts will evolve from generation to generation.

From Christian doctrine alone, the historian cannot explain anything about the institutions that Christians have chosen to create. Instead, the historian must develop an intricate model of the kind of society in which the institution was born. He must conjecture the degree to which the society is dynamic or static, whether the memes of the society are largely rational or anti rational, whether the Christian memes are largely rational or anti rational, what the relationship is between incumbent Christian institutions and the rest of society, the level of knowledge and wealth of the society, and, of course, the processes that characterize the Christian institution in question. Only in light of this set of interconnected cultural theories does a historian have any hope of explaining the historical role of a particular Christian institution.

Why should the historian care whether the society under scrutiny is dynamic or static, whether the relevant memes are rational or anti rational? Because explaining the

evolution of ideas is not merely a matter of judging the truth content embedded in them. On the contrary, the institutions of some societies are such that the *worse* an idea, the more effective it will be at spreading throughout the population.

Consider the historian who argues that Darwin's theory of evolution spread throughout the scientific community of the 19th century West because it is true. First of all, fallibilism tells us that Darwinian evolution is not true per se—it may be our best explanation of the regularities of the biosphere, but it necessarily contains errors and gaps that a future, deeper theory will resolve (in fact, the neo-Darwinian synthesis and constructor theory of life have since done precisely that). In other words, Darwinian evolution contains more knowledge than its competitors (such as Creationism), but it does not contain all knowledge for all-time. Secondly, new ideas that contain more knowledge than their rivals have an easier time promulgating in a dynamic society than in a static society. So, for the historian to explain the rapid adoption of Darwinian evolution, he must refer to the scientific community's cultural attributes—openness to new ideas, a desire for progress, and a robust tradition of criticism.

A tradition of criticism is one of the most important institutions that a culture can adopt. As we have seen, knowledge grows by creative conjecture and criticism. For a culture to consistently replace worse ideas with better ones, it cannot just criticize incumbent ideas once. It must have the capacity to do so continuously and across evermore domains of society.

Ancient Athens' tradition of criticism may have begun with philosopher Thales' Milesian School. Pupils regularly criticized the ideas of their masters in an effort to improve them. Thales's own student, Anaximander, rejected some of his teacher's ideas in favor of his own. It may well be that the Milesian School was the first institution built precisely for the purpose of argument in pursuit of the truth.

Thinkers of the Enlightenment era similarly established a tradition of criticism that, unlike Ancient Athens', survives to this day. One of the most important roles that a historian can serve is to explain why some (indeed, probably most) traditions of criticism have gone extinct, while others survive. What attributes must a tradition of criticism possess to survive threats to its existence? Does the answer depend on how dynamic or static the broader culture is? Does the answer depend on how dynamic or static the tradition of criticism is?

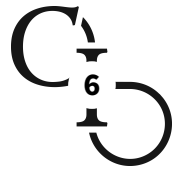
As we have seen, bad epistemology can cause historians to mistakenly castigate the West as the ultimate evil and venerate static societies as innocent, pure, or noble. Epistemology is intimately connected to morality—if progress in every sense depends on the growth of knowledge and the correction of errors, then it is morally wrong to degrade any institution that catalyzes such a process.

Some historians are conspiracists, chalking up every war and calamity to shadowy interests who puppeteer society and cause chaos for self-interested reasons. Ironically, this means that they neglect the *actual* reasons why individuals make their choices: conspiratorial historians fail to explain *why* a government might go to war, *why* it might implement a mistaken policy. All people act on reasons, and the explanation of those

reasons must cohere with the *rest* of our best explanations from history, economics, institutional dynamics, and memetics (culture).

The historian who does not know economics is liable to attribute economic booms, busts, poverty, and wealth to the wrong causal factors.

We are actively seeking a historian Fellow who shares our appreciation for how Deutschian critical rationalism can facilitate good history. In the meantime, Logan is developing a course for Conjecture University, *Civilization*, that will touch on many related ideas.



Conjecture Institute's Four Branches

1. ***Conjecture University*** includes original research in fundamental physics and mathematics as conducted by select Fellows and our Senior Scientist. It also includes online courses, whose content and structure we are currently developing.

    a) Fellow Samuel Hagh Shenas is currently producing a series of written and video lectures about the history and philosophical significance of the principle of locality for a course titled *The Principle of Locality*.

    b) Fellow Maxime Desalle explains why the Everettian interpretation makes sense in his course, titled *Taking Schrödinger Seriously*. Designed for a lay audience but without watering down the ideas, Maxime's course investigates the central equation in quantum mechanics and explains the implications for taking it seriously as a reflection of what reality is actually like.

    c) President and Cofounder Logan Chipkin is currently developing courses on epistemology, constructor theory, economics, and civilization.

    d) Each course module will not be created in isolation. Rather, their subject matter will be chosen deliberately so as to fit into Conjecture University's broader library of video educational content. Conjecture University's library will ultimately consist of interrelated modules that can be consumed one-by-one or 'in order', as some modules can serve as precursors to others.

2. ***Conjecture Press*** is our in-house book publishing arm. We have already published four books and have one more in production. Many of our Fellows are also writing books that we will publish.

a) *The Sovereign Child: How a Forgotten Philosophy Can Liberate Kids and Their Parents*, by Conjecture Institute Cofounder Aaron Stupple with President Logan Chipkin. Could it really be okay to let kids eat whatever they want? Sleep whenever they want? Watch whatever they want? If kids are completely free to make their own choices, won't they develop damaging habits that will haunt them into adulthood? Surely parents have a duty to set a few limits. But what if a forgotten philosophy from the 20th century explains why this conventional wisdom is wrong? In The Sovereign Child, Aaron Stupple carries the torch of Taking Children Seriously, a parenting movement whose cornerstone is the idea that children's reasons, desires, emotions, and creativity all work precisely the same way that those of adults do—in short, that children are people.

b) *Lords of the Cosmos: From Stasis to Stars*, by Conjecture Institute Fellow Arjun Khemani and Logan Chipkin. Most of human history has been unremarkable and static. But Western civilization is different, sustaining rapid progress for generations. And we're just getting started. *Lords of the Cosmos* views the story of humanity through the lens of our most profound theories of progress, addressing such questions as: What sparks progress? What does progress entail? What is the role of human progress in the cosmic scheme of things? *Lords of the Cosmos* illustrates how humanity is the most powerful force in the universe, capable of creating any object the laws of physics allow. Authors Arjun Khemani and Logan Chipkin argue that suffering is intimately related to staticity, while creativity provides the only means to end suffering.

c) *Bold Conjectures, Volume I: Select Interviews of David Deutsch*. David's two books, *The Fabric of Reality* and *The Beginning of Infinity*, offer a deep and coherent worldview that has improved on humanity's ideas in physics, epistemology, morality, aesthetics, and other fundamental domains of knowledge. He's been interviewed over one-hundred times during his career, giving his readers hours of additional content that elucidates and expands upon the ideas in his books, in addition to ideas far afield from his writings. David's interviews provide more than enough content for a book—and that is precisely what we've done. *Bold Conjectures, Volume I: Select Interviews of David Deutsch* is a compilation of over a dozen interviews of David Deutsch.

d) *The Farthest Reaches: Why People Are the Most Important Entities in the Universe*, by Conjecture Institute Ambassador Brett Hall. Brett Hall expounds on the concept of *explanatory universality*, as first explained by physicist and philosopher David Deutsch in his book, *The Beginning of Infinity.* It is this characteristic that grants people their primacy in the cosmos. This ability to explain anything that can be explained, to understand anything that can be understood, is more significant than a star's

titanic gravitational pull, a gene's ability to replicate itself, or a computer's ability to execute calculations at the speed of electrons. Beyond *explaining* this idea, Brett *applies* the concept of explanatory universality to a number of contentious debates in society, such as: school, IQ, multiculturalism, mental illness, evolutionary psychology, immigration, and many more.

e) *Bold Conjectures, Volume II: Essays Across Physics* consists of about a dozen original essays about concepts in physics that either the public does not appreciate as much as physicists do, or else cutting-edge ideas from various subfields. Expected release time is Q4, 2026.

f) The next two entries in the Bold Conjectures series will consist of previously published essays by David Deutsch and a compendium of essays on the topic of civilization.

g) We expect to publish at least some books by Fellows such as Carlos (AGI), Tom (Wonderism), Maria (physics thought experiments), Eric (Darwin and Popper), and Ray (bringing the Enlightenment to the modern world).

3. ***Conjecture Studios*** is our media arm, which includes documentaries, science fiction films, and short videos that explain ideas with the help of animations.

    a) Our podcast has consisted of Brett Hall interviewing Advisors, Fellows, and Cofounders, but we will be expanding the roster of interviewees in 2026 to thinkers and builders who work on problems that overlap with our own. Logan will also conduct interviews with many contributors to *Bold Conjectures, Volume II: Essays Across Physics* that will serve as an audio companion to the written compendium.

    b) Arjun's documentary was just the first of many long-form films we'd like to produce. We have ideas for a Taking Children Seriously documentary, as well as a fictional animated film that tells the story of the evolution of a tradition of criticism.

    c) We are producing and will eventually host Dimitri Vallein's short film, *The Day I Met You*, in 2026.

    d) Fellow Jaber Hassoun has begun to create short- and medium-length video essays that we will host.

4. ***Conjecture Forum*** is our events arm.

    a) In the early fall, we hosted *Rat Fest 2025*, our fourth ideas festival, this one attended by over 75 people from all over the world. Attendees raved about the special opportunity to meet in-person and build relationships with enthusiasts who are otherwise reachable only online. We will continue to host this event annually (rebranded as *Conjecture Con* henceforth).

b) In June 2026, we will host an online physics conference focused on constructor theory, the Heisenberg picture of quantum mechanics, physics without time, Everettian quantum theory, and quantum gravity.

c) In the fall of 2026, we will host an online conference whose theme will be 'Defending the West'.

Conjecture Institute's Advisors

Judea Pearl

Judea Pearl is a computer scientist and philosopher whose work has transformed the study of causality and artificial intelligence. He is Professor of Computer Science and Director of the Cognitive Systems Laboratory at UCLA. Best known for developing Bayesian networks and the do-calculus, Judea has provided the foundations for modern causal inference, reshaping how science, statistics, and AI address cause-and-effect questions. His books, including Causality and The Book of Why, have influenced fields ranging from epidemiology to economics. Among his many honors, he has received the A.M. Turing Award, the Benjamin Franklin Medal, and the Harvey Prize.

Daniel Hannan

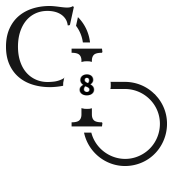
Daniel Hannan, Lord Hannan of Kingsclere, is an author and columnist, and is President of the Institute for Free Trade. He teaches at the University of Buckingham and the University of Francisco Marroquín in Guatemala. He has written nine books, including the Sunday Times bestseller How We Invented Freedom. He sat as a Conservative MEP for 21 years, and was a founder of Vote Leave. He served on the UK Board of Trade from 2019 to 2024. He writes regular columns for, among others, The Daily Mail, The Sunday Telegraph and The Washington Examiner.

Peter Boghossian

Peter Boghossian serves as the Executive Director of the National Progress Alliance and Founding Faculty Advisor at the University of Austin. Drawing on over 25 years of

teaching experience, he specializes in the Socratic method, scientific skepticism, and critical thinking. His doctoral research developed innovative methods to enhance moral reasoning and reduce recidivism among prison inmates. His last book, How to Have Impossible Conversations, has sold over 100,000 copies and been translated into ten languages.

Peter engages a wide audience online, with over 236,000 YouTube subscribers, 27 million views, and 331,000 followers on X. His work has been featured in publications such as The New York Times, The Wall Street Journal, and Scientific American, and he's been on top podcasts including The Joe Rogan Experience and BBC's HARDtalk. Additionally, he has made significant contributions to peer-reviewed literature.



Conjecture Institute's Leadership Team

Logan Chipkin

Logan Chipkin is a writer and editor in Philadelphia. He has written articles about fundamental physics, economics, history, and Popperian philosophy for outlets such as *Gizmodo*, *Physics World*, *Quillette*, *History Magazine*, *The Libertarian Institute*, *The Pennsylvania Gazette*, and *Bitcoin Magazine*, and he has written a fantasy novel called Windfall. He has also edited a book written by a prominent physicist and has collaborated with other physicists to communicate their ideas to a general audience.

Aaron Stupple

Aaron Stupple is a practicing physician and father of five in Western Massachusetts. He has been promoting critical rationalism and the work of Karl Popper and David Deutsch since 2019 in the form of online community building, a web magazine, and Rat Fest, the annual in-person conference in Philadelphia. Currently, Aaron wrote a book on Taking Children Seriously, which is the application of critical rationalism to parenting.

David Kedmey

David Kedmey is the president and cofounder of a financial technology company that helps professionals derive trading insights from historical data. His interests span AGI, biomorphs, and pedagogical software games. David has a particular passion for mentoring and supporting those working to promote and develop the ideas of Karl Popper and David Deutsch.