

57 Ideas on Cognitive Infrastructures: Synthetic Intelligence in the Wild Benjamin Bratton

As AI becomes both more general and more foundational, it shouldn't be seen as a disembodied virtual brain. It is a real, material force. AI is embedded into the active, decision-making processes of real world systems. As AI becomes infrastructural, infrastructures become intelligent. Natural intelligence emerges at environmental scale and in the interactions of multiple agents. It is located not only in brains but in landscapes. Similarly, artificial intelligence is not contained within single artificial minds but extends throughout the networks of planetary computation: it is baked into industrial processes; it generates images and text; it coordinates circulation in cities; it senses, models, and acts in the wild.

This represents an infrastructuralization of AI, but also a "making cognitive" of both new and legacy infrastructures. These are capable of responding to us, to the world, and to each other in ways we recognize as embedded and networked cognition. Cognitive Infrastructures are forming, framing, and evolving a new ecology of planetary intelligence.

The lecture 57 Ideas on Cognitive Infrastructures: Synthetic Intelligence in the Wild explores this work as part of Antikythera's Summer Studio on Cognitive Infrastructures, based at Central Saint Martins, University of the Arts London and hosted by CSM MA Narrative Environments and Digital Innovation. This lecture builds on the *After Alignment* keynote, filmed in 2023.

Benjamin Bratton is Professor of Philosophy of Technology and Speculative Design at the University of California, San Diego. Through the lens of planetary computation, his work establishes new philosophical frameworks for interpreting the past, present and future co-evolution of life, culture, and technology. He is Director of Antikythera, a think-tank researching the future of planetary computation based at the Berggruen Institute. He is the author of numerous books including *The Stack: On Software and Sovereignty*. The tenth anniversary edition will be published by MIT Press in 2026.

EDITORIAL

This talk, "Cognitive Infrastructures: Synthetic Intelligence in the Wild," presents 57 interconnected ideas framing AI not as a singular entity but as an evolving environmental phenomenon. It was filmed at Central Saint Martins at King's Cross, London on the occasion of Antikythera's Cognitive Infrastructures studio and features key ideas from this work.

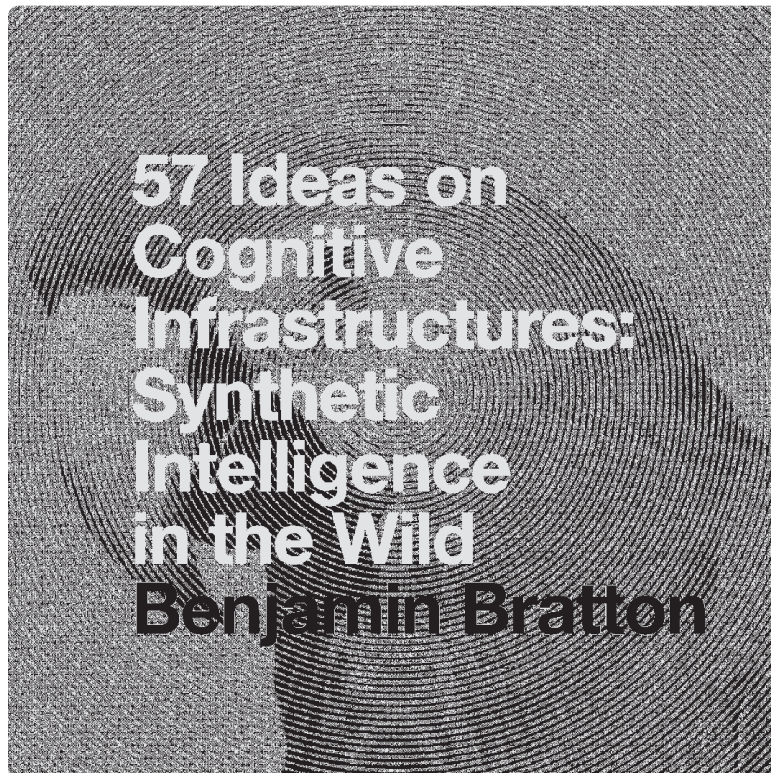
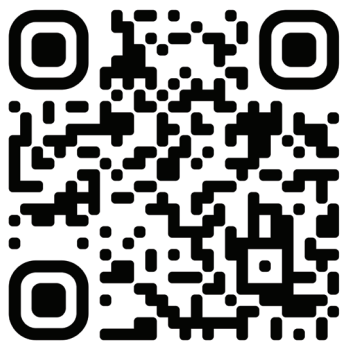
The talk positions computation as both an instrumental tool and an existential technology that reshapes our understanding of ourselves and the planet. The core concept is Planetary Computation, where AI is part of a cognitive infrastructure that both models and transforms Earth.

Bratton argues AI is evolving literally, extending intelligence beyond biology into the "lithosphere" (mineral intelligence). This evolution follows a pattern: autopoiesis (self-making) leads to allopoiesis/artificialisation (making external tools), selecting for intelligence, then symbolic language, and now AI. AI recursively accelerates this entire cycle, acting as a scaffold for future, unknown developments.

The talk critiques anthropocentric models of AI (like the Turing Test), advocating for understanding AI "in the wild"—distributed, embodied within environments like cities or ecosystems, interacting in complex ways. This involves cognitive infrastructures where artificial intelligence becomes infrastructural, and infrastructures become cognitive, interacting with natural forms of automation (ecological processes).

Human-AI interaction (HAIID) is explored through concepts like productive dis-alignment (questioning if perfect alignment is always best), the inverse uncanny valley (discomfort from AI's perspective on us), and reflexive, often asymmetric, mind-modeling between humans and AI agents in various ratios. Interfaces, broadly defined, are where this coevolution occurs.

Generative AI is discussed not just as media automation but as modeling the "manifold" of culture, leading to new political dynamics like "derepresentation" or manipulating the AI's model of reality ("Potemkin politics"). The talk challenges the traditional definition of language, suggesting AI expands it towards calculation and tokenizable information, potentially evolving languages beyond human control.



PART ONE: INTRODUCTION

00:00

Part One: Introduction

Thank you Stephanie for this wonderful introduction, and thank you all for coming. I was looking out at all the friends and faces that I know. It's a really wonderful group that came tonight, so thank you. And we'll have some time to hang out in the bar upstairs afterwards. Other thank you's before I begin, are to Central Saint Martins for hosting not only this event, but the Cognitive Infrastructures Studio; to the whole Antikythera team whose tireless work has put all this together, to our Studio Researchers who decided to spend a month with us, working through some of these ideas; our Affiliate Researchers who flew in from all over the world to think with us; and our many partners, including and especially the Berggruen Institute, some of whom are here with us this evening.

Let me give a little bit of an introduction to what this is all about, for anyone who is new to our little conspiracy, and to start with the definition of Antikythera itself (a bit of a tongue twister). The term and the name Antikythera comes from this little device, the Antikythera mechanism, which dates from the second century BCE. It was discovered in 1901 off the island of Antikythera from whose name it bears. Though we don't really know what it was called in its time.

It was, so the story goes, the first computer. But it wasn't just a computer. It was also an astronomical device. And so the idea that computing begins in the orientation of intelligence in relation to its planetary condition, is one that we find quite appealing, one that extends through history and provides an interesting and important filter to the long story of what computation is.

Speaking of which, let me begin with a little invitational thought experiment—to imagine the Blue Marble image taken by Apollo astronauts in the 1960s not as a still image, but rather as a kind of movie showing all 4.8 billion years of Earth spinning around super fast, mostly going from taking a long time to turn from red to blue and green, mostly everything that happens, you're going to see in the last couple seconds. And at the very end what you see is this little blue marble sprout, an exoskeleton across the satellites and sensors, and fiber-optic cables transversing its oceans of cities and highways. But most importantly this outer layer. One of the lessons to draw from this little parable is that it's not just that things evolve on a planet, but planets themselves evolve.

The Great Oxidation Event, in which the appearance of oxygen emitting plants causing a massive die off 2.3 million years ago produced the atmosphere, which in turn made the evolution of yet more complex life possible in the first place. A form of complex life that eventually built that exoskeleton. Intelligence is also then not just something that happens on a planet. It's something that a planet does. And some planets are better at it than others, our's being pretty good. But that exoskeleton didn't just appear, it didn't evolve in a purely random way. Obviously humans did it. But humans are part of the biosphere, and so as biospheres create technospheres, we're also sensing how technospheres in turn create biospheres. In a way, that's really what tonight's talk will ultimately be about.

Stephanie mentioned this term, planetary computation. Those of you who are familiar with The Stack will know a little bit about this. Part of the idea is that we need to think of computation not only in terms of mathematics and calculation, but also as a kind of cognitive infrastructure. Something that reveals the planetary through things like climate science, but also is reproducing and is becoming a medium for the transformation of planetary systems. It itself, with the emergence of AI, is going through a kind of fundamental transformation. And how we produce information, calculate information, distribute information, and in essence, grow things with information is changing the entire architecture of planetary computation incredibly quickly. It may very well be that the first phase of planetary computation, the one we built over the last 50 years, was really just a precursor to the real story of what comes next.

Before I really begin, let me just give you a little bit more ideas to think with. Another one is that for us, computation, obviously a term that could mean lots of different things, but that computation was discovered as much as it was invented. Computation doesn't explain everything, but that it is part of the natural world, and what we produce with these silly little appliances is really artificial computation— a low dimensional approximation of a natural process, that is far more efficient and powerful than what we have built today. Hold this in mind for a moment.

Every story needs heroes and villains. One of the heroes of our story is the Polish science fiction author Stanislaw Lem, whose work you probably most know from the novel *Solaris* and then the Tarkovsky film afterwards. He wrote almost entirely fiction, but there was a really important book of nonfiction that he wrote in the early 1960s called *Summa Technologiae*. One of the key lines, it's not a major theme, but one of the key aspects of the book was pointed out to us some time ago by Bogna Konior, who's here with us tonight: Lem makes this essential distinction between what he calls instrumental technologies and existential technologies. This is what I want you to hold on to.

Instrumental technologies are those, so Lem says, that their main impact on society is how they operate as tools. So bulldozers move dirt and so you can make cities faster. Fine.

There are also existential technologies. These are technologies which, when used properly, also change the way in which we understand how the world works and how the universe works. Microscopes, telescopes—they let us see things very small, very far away, but they also transform our understanding of what life is. They enable the heliocentric cosmologies which transform our understanding of what we are, where we are, and when we are. For us, computation is both an instrumental and existential technology. And for tonight's discussion I want to put a little bit of emphasis on the aspect of it as an existential technology: how computation is transforming our understanding of who, what, when, where and how we are.

Let me give you a little bit of ground rules for the talk tonight. There are lots of AI talks. You've probably exhausted by the very idea of another talk on AI. This one is hopefully a little bit different. Most of the talks you might have sat through are trying to present some kind of grand unified theory of what AI is, as if it were possible even to ask the right questions and get the right answers at this early phase in the evolution of AI. This talk is really different. It's much more bottom up. These are some ideas that we are working with and thinking through and have come up with in the first half of the Cognitive Infrastructure Studio that we'd like to share with you. You can think of it as a bit of an invitation into the studio itself to join us in that conversation. For me, it's really a kind of report on things I've been thinking about for the last three months or so. It's a work in progress. So, yes, not another AI talk.

On the Cognitive Infrastructure Studio let me paint a little bit further how this is set up. A part of the way in which Antikythera works is by working with, as Stephanie had indicated, a whole range of researchers and thinkers and makers and doers from lots of different disciplines who seem to have some kind of common spirit with the things that we're doing. But one of the things that's been very important to our thinking and that goes all the way back to the Strelka days is thinking and making in conjunction together. These studios that we run, which are quite interdisciplinary, are a moment of really intense acceleration of some of our thoughts.

Our studio researchers are here with us this evening. They come from a number of different disciplines and institutions and walks of life, and I'm sure they'll be eager to tell you a little bit more about what we've been doing. Our Affiliate Researchers: Blaise and Stanley and Sarah and Thomas. These are people with whom we've been thinking for years. And in that conversation and dialogue with them the very possibility of the questions that we are asking in the studio came to form. The opportunity to work with them and to try to play with these questions we're very grateful for. Of course, all supported and executed by quite a growing team of Antikytherans—the Kytherans—that's a whole other story.

The relationship then, is clearly between the design and philosophy, the way we're thinking about it. The way I see it is this: it's not just a kind of thinking through making, but that at different points in history there's a different relationship between our ideas and what it's possible to do. For certain periods of history our ideas of what we would like to do are way ahead of our actual capabilities. So when the Russian cosmonauts were dreaming of spaceflight in the 19th century the ideas were ahead of the technological capacity.

Today, I think, something different is at work. I think our technologies and what is possible to do is ahead of our ideas. We don't have the concepts, the language, the terminology, the frames of reference to even understand and articulate what is happening right in front of us, let alone to orient it. And so for us, the philosophy is not projected onto the technology. It's derived from a direct engagement with it.

AI, this particular moment, this particular topic for us is incredibly exciting. There is, of course, a long standing culture war between the thinkers and the makers, between humanities and engineering. But when what we're talking about is making "thinking", the thinkers are the makers and the makers are the thinkers.

What we'll be presenting then—you could think of as puzzle pieces—concepts as puzzle pieces. You might imagine the 57 ideas that I'm going to present to you tonight as bits of a puzzle piece that can fit together into lots of different kinds of images. Some may be more interesting than others, but there may not be a final resolution to these images that kind of explains everything. Part of what I'd like to share with you is how we're playing with them and thinking of them not as potential ending points, but really as starting points.

With that said, no idea is really separate. Any idea that is worth thinking through

replicates itself, reproduces itself. It becomes a scaffold for yet more complex ideas, just like life scaffolds itself. And for me, this is part of the ethics of philosophy—that it's never done, that the best thing you can hope to build is something that later on becomes part of something else. You build something that others can build with later on. So this is again our invitation.

PART THREE: THE EVOLUTION OF THE ARTIFICIALIZATION OF INTELLIGENCE

14:04

Part Three: The Evolution of
the Artificialization of Intelligence

I've mentioned this term "evolution of artificial intelligence", and I actually don't really mean it metaphorically. I mean it quite literally. I think machine intelligence is evolving, and evolving through processes that are roughly similar to how life evolved. And in many ways, as Sara has made clear to us, AI really is life.

One way to think about it, going back to our Blue Marble picture, is that we've had many millions of years of animal intelligence, which became *Homo Sapiens* intelligence, we've had many millions of years of vegetal intelligence, and now we have mineral intelligence. That the fire apes, that's us, have managed to fold little bits of rocks and metal into particular, intricate shapes and run electricity through them. And now the lithosphere is able to perform feats that, until very recently, only primates had been able to do. This is big news. The substrate (this will be part of my attempt at a conclusion this evening) the substrate of complex intelligence—that's us—now includes both the biosphere and the lithosphere. And it's not a zero sum situation. The question that we are beginning to be able to ask is: how did those integrate together in such a way that they become mutually reinforcing, not mutually antagonistic?

Another question to consider is what do we mean by "artificial" in artificial intelligence? And this is a topic of a long research project that we've begun. I'm not going to try to unpack it all here, but one of the ways (from information theory) that we may think about the artificial is as anomalous regularity. Anomalous regularity.

Everyone in this room, I hope, can perceive which of the trees in this Japanese forest were planted by people, and which of them evolved in place. And the reason you're able to do so is because you perceive a kind of anomalous regularity. Good.

There's lots of ways of unpacking and making use and repurposing this term 'artificial'. And, I'm going to go through a different kind of one. The argument I want to make to you—it might seem circular, but it's because the argument is about something that is circular—is that in some respects we can see AI as the artificialization of artificialization itself. Let me try to tell that story.

A few days ago, I made this diagram with Nicolay and Sara and some others. I'm not going to walk you through this particular version, but when I was a kid, I drew constantly, I was drawing basically all day long. For me, learning how to think and learning how to draw were kind of the same process. Now I'm a writer, but in many ways I think of writing as a kind of drawing. So with Nicolay's help we developed a much more legible version of this. I'd love to do a talk where I really just draw through the ideas in time, but that's a different process.

Once upon a time, let me tell you another kind of parable. You can decide how much of this is fact or fiction. There is evolution. Evolution is a process that I'm convinced, and Sara's convinced me, begins not with biology, but with chemistry at the very least, and that it's a functional dynamic of selection. Eventually, it arrives at something we might call life, that it provides the scaffolds for something that emerges as life which can take many forms.

But one of the key ideas is autopoiesis. But really it's that autopoiesis allows forms of chemistry to become life, because being alive or being life allows them to capture more energy, matter and information which allows them to reproduce themselves,

to persist. This is autopoiesis; the ability to take the part of the outside world and use it to remake yourself over and over.

To get really good at autopoiesis it's also useful to figure out ways to make use of the outside world to capture information, energy, and matter that produce things external to you. This is the basic dynamic of artificialization—allopoiesis; building things that allow you to capture energy, information, matter extrinsic to yourself and this is how you get really good at life. Each one of these is a scaffold for the other.

To get really good at artificialization, this tends to select for intelligence, complex abstraction to figure out ways to do artificialization really well, to have a hylomorphic projection, to produce complex models of what might be counterfactual models of what might be, and then instantiate them in the world.

To get really good at intelligence and to instantiate those counterfactual models and indeed communicate them among a lot of other people to coordinate with you you need a kind of symbolic language. So intelligence in turn selects for symbolic languages, written encoded languages in various combinations, one of these building on the other. Though it's also true that they kind of recursively impact each other, that once you get symbolic language, this changes how intelligence works, which changes how artificialization works. And so it doesn't only go one direction, though it appears in a certain order.

And so artificial intelligence, in this evolutionary cycle, is making use of the symbolic languages. It's trained on its large language models, and whatever comes next are at present trained on the symbolic languages of what this makes possible. As I said, some of these fold back on each other. I should also say that something else, of course, will come next.

But we're also seeing that artificial intelligence is already and will continue to recursively transform how symbolic language, written encoded language, and all its forms will be produced and structured and developed, which will in turn impact how intelligence works. It goes both ways, but the really important thing in many respects is that it will recursively impact how artificialization works, how it is that it's possible to transform the world, capture energy, information, management, to make things that are extrinsic to machine intelligence, in ways that are considerably more complex and intricate, than anything that has happened in the past.

What we want to think about is the location of AI in the evolutionary arc, but that doesn't just mean in the past, it also means in the present. And to think about AI in the present, through evolutionary dynamics is really the basis of the cognitive infrastructures idea, but also in the future, because this isn't ending. AI is not the last thing. Just as intelligence was a scaffold for symbolic language, which was a scaffold for AI, AI is a scaffold for something else, which is a scaffold for something else, which is a scaffold for something else, and so on and so on. What we're building is a scaffold for something unforeseeable, but something that has a deep future, as well as something that's based on deep time.

So, Cognitive Infrastructures. As said, it has to do with an understanding of the evolution of AI in the present or evolutionary dynamics of AI in the present. Let me explain a little bit of what we mean by cognitive infrastructures, and you'll get a better sense of the kind of work that we're doing.

Not only does philosophy and design have this kind of cyclical relationship to each other, but for our work, throughout history, the philosophy of AI and the actual engineering of AI have co-evolved in a kind of twisting double helix where thought experiments generate real technologies which generate thought experiments which generate real technologies. And with that long term future in mind, that dynamic can and should continue.

One of the key ideas for cognitive infrastructure is thinking about AI as an external force. That AI would not only be something that we live with, it will be and already is, in many respects, something that we live in. To live in as part of the environment. And something that as an actor and an agent and an object, within that environment. The model of AI that cognitive infrastructures implies is different in many respects than conceptual models of AI that we have come to rely on.

The Turing test, or the Voight-Kampff empathy test is based on the idea that what Turing proposed as a kind of sufficient condition of intelligence—that is, if the AI can perform thinking the way that we think, that we think—then there's something interesting going on there, and we almost don't even need to decide whether or not this constitutes an intelligence like us or not, which has then become over in popular culture a necessary condition of intelligence—that is, unless it can establish this high degree of granular anthropomorphism, it is disqualified.

But the real problem for us is this idea that intelligence exists in the mind of a single organism in the first place, and that the mirroring of the mind of a single organism in an AI is how to identify and encounter the AI in the first place. Which has all kinds of downstream effects in the way in which we are framing the question of what kinds of machine intelligence we want to build, what the interfaces and relationships between human societies, human intelligence, machine intelligence might be like, the training to expand the amount of anthropomorphism in ways that may be useful in many respects. And also might be dangerous in others.

There's many other precedent models and metaphors, allegories of AI to draw from. Certainly not all of them from the West. One of my favorites is the Alpha 60 urban AI from Jean-Luc Godard's *Alphaville*, which was represented in this film as just this glowing dot, which Kubrick sort of borrowed for 2001 a few years later. But, what Godard was shooting when he filmed this was just the elevator button in his apartment building, which we took to be the AI.

Speaking of *Lem*, in *Solaris*, the entire ocean of the planet is intelligent. It's a kind of massively distributed brain, and because it's embodied so differently it has a bit of difficulty communicating with the cosmonauts, if you know the story. But this is an idea of artificial intelligence that is not about it being contained in a single organism, but it's something that exists as part of what a planet is doing, as part of an expanded environment. And this is the thread that we're pulling on for the Cognitive Infrastructures Studio.

That is because natural intelligence evolved in open worlds in the past, the presumption is that we should look for ways in which machine intelligence will evolve in the present and future through open worlds as well. This also means that its substrates of intelligence may be quite diverse. They don't necessarily need to be human brain tissue or silicon. They may take many different forms. This is not my refrigerator, by the way.

Another way of putting this is instead of the model of AI as a kind of brain in a box, we prefer to start with the question of something more like AI in the wild. Something that is interacting with the world in lots of different, strange and unpredictable ways. There's always a lot of discussion about how AI may or may not be embodied, and how its embodiment or disembodiment structures how it's possible for it to think, because of how the embodiment structures how it's possible for it to be in the world, to perceive the world. And obviously what we have in the past called robotics and may soon just call physicalized AI, is obviously a way in which AI is increasingly embodied in the world, learning through its tumultuous encounters with real world situations just as baby animals do. If we take the massively distributed model more seriously, we may see that, in many ways, the AI is embodied as something like the whole city, or the entire artificial landscape. Forget the word city. The entire artificial and natural landscape combined is a form of physical embodiment for AI, even if it doesn't look like the tetrahedral body plan or any kind of animal that we may be familiar with.

This connects with one of the other ideas that we've been working through, going back to the days of *The Terraforming at Strelka*, one that's been picked up in different ways in a lot of Stephanie's work as well— it's what we call the ecological theory of automation. This requires a bit of explanation, but long story short, the idea is that automation is also already part of ecological process, part of natural process. That automation is part of how living systems work. That from the whole potential chaos of contingency, of how agents might interact with each other at different points within that system, there's a kind of congealing of something that becomes stabilized and becomes stabilized, it is able to repeat functions over and over and over, input and output functions. And because it repeats them over and over, it becomes predictable. It becomes something around which other things can grow. And those trophic cascades of things affecting other things, affecting other things in regular, even homeostatic kinds of cycles is part of how ecologies work. So what we see as you walk into a complex factory is second order automation, what you see there is a low dimensional approximation of natural automation that is the world itself.

And so why is this important for cognitive infrastructures? Well, if you're talking about AI in outside environmental physical systems, the question of automation would come immediately to mind. But we want to ask this sort of a first question of how are the ways in which that artificial automation might interact with, and interface with, and connect to, and learn from the natural forms of automation, those trophic cascades in ways that are not adversarial, but in fact, symbiotic.

Important thing to keep in mind is that as opposed to *Skynet* or AI digital agent on your phone, AI, as natural intelligence, can be part of every single agent in a long, complex chain of causality in a big chain of systems, every one of these little agents could be intelligent in some way. So too for artificial intelligence. Any way in which you might imagine the tumbling of the dominoes, that is the cascade of a causality through a complex system. AI isn't just the first domino pusher, and it's not only the thing that happens at the end, it's something that can be in every part of it as the whole thing goes along in lots of different ways. That just as you have lots of hierarchical speciation in nature of little tiny things with very specific kinds of minds and medium things with medium minds and big things with big minds and so forth, different combinations of these, that the ecosystem and the technodiversity of artificial life is something that is generative of that complexity and also an outcome of it as well.

Where do we fit into this? Well, potentially lots of ways. The way of thinking of AI as a kind of something that's in competition with humans I think is a bit of a traumatic response which I'll talk about in a moment. I think that in many ways it's quite possible that the emergence of AI is the thing that will make the long term evolution of *Homo Sapiens*, more possible than it would have been otherwise.

But as you walk around the city now, and I want to try to convince you that the scenarios that I'm talking about are not something that's going to happen in the future. This is not like "in the year 2100 there'll be flying cars" kind of story. This is a way of describing what's happening right now outside these four walls; that as you walk around the city, with all of the neurons in your brain firing, looking around and interacting with things and looking at things in different kinds of ways, you're able to do so intelligently because of the multiplication of synapses in your brain. It's important to remember that if you have a more or less recent phone, that you have more neurons in your hand than you have in your brain, that the distributed cognition, as it's called, in the embodied cognition is something that is not hypothetically about an amalgamation of human intelligence and machine intelligence. That amalgamation is exactly how the city sees you. And sees you as something that is capable of both modalities of intelligence at the same time. Maybe another way of naming it that might unlock it for others is that the urban fabric is a stand-in for any kind of massively complex artificial environment is a "landscape". It's a landscape for obviously lots of organic and natural intelligence, but also a landscape for inorganic distributed intelligence. That's the 'where' the cognitive infrastructures really are.

PART FOUR: COGNITIVE INFRASTRUCTURES

Let me summarize this a little bit before I move on to some of the more specific ideas that we've been kicking around with as well.

Natural intelligence emerges at an environmental scale through the interactions of multiple agents of varying nested complexity. For cognitive infrastructures the same is true of artificialized intelligence.

34:18

Part Four: Cognitive Infrastructures

As artificial intelligence becomes infrastructural, infrastructures concurrently become more cognitive. They are capable of responding to us, to the world, and perhaps most importantly, to each other.

As artificialization becomes both more general and more foundational, it shouldn't be seen only as a disembodied virtual brain, because sometimes it is, but also as a real material force that is not the opposite of life; it is a kind of life.

Perhaps the most critical interactions will unfold between different AI's, without a lot of human interference. The sensing and modeling and acting in the wild and so forth. Cognitive intelligences are infrastructures, forming and framing and evolving a new ecology of planetary intelligence.

PART FIVE: PRODUCTIVE DISALIGNMENT

35:47

Part Five: Productive Disalignment

This is not our conclusion. It's our starting point.

It's the heuristic of questions that we're asking and trying to find out what we can build and make with these questions. Some of these questions, that bigger question, has led us in particular ways. I want to share a little bit some of the pathways that we've been going down, because maybe you have different ideas of things that we could build with it.

One of the key ideas is what we call productive disalignment, which I'll define for you. This is very meta, this is my talk here last year, it was a talk called After Alignment. This talk took place right after that big letter that everyone signed that was about how we should push the big red button and stop AI, and I was very suspicious of the whole thing. You can watch this online.

One of the ideas that was in that talk was an idea called Reflectionism. Reflectionism is sort of the paradox within AI discourse. There's one part of AI discourse that says: AI is just a direct manifestation of the biases and injustices of our political economy. That it is an extrusion of society as it exists now, that it is fundamentally culturally determined, economically determined, and that regardless of anything we do it is always already a reflection of us. There's another side within AI discourse which says that the real problem with AI is that it's not like us, but needs to be made like us, which is a simple way of describing what alignment might be like. That it should be like us, but it isn't. This paradox between 'it is like us, and that's the problem.' And, 'it should be like us but isn't,' both of these depend or orientate or revolve around the idea that in some way, human intelligence and machine intelligence are reflections of each other. Maybe they are, maybe they aren't. But I don't think either one of these particular two directions actually gets us to all the ways in which that reflection is actually real and interesting, and something that is not always to be, that necessarily needs to be critiqued or invented.

Let me tell you a little story about this. And I'm going to read from this, so I get it right for you. At an OpenAI retreat not long ago, Ilya Sutskever, until recently the company's chief scientist, commissioned a local artist to build a wooden effigy representing unaligned AI. Ilya then set it on fire to symbolize 'OpenAI's commitment to its founding principles'. This curious ceremony was perhaps meant to preemptively cleanse the company's work from the specter of artificial intelligence that is not

directly expressive of human values. Just a few months later, that topic became an existential crisis for the company and its board when CEO Sam Altman was betrayed by one of his disciples, crucified and then resurrected three days later. Was this alignment with human values? And if not, what was going on?

Also consider at the end of that same last year, Fei-Fei Li, the director of Stanford's Human-Centered AI Institute, published, 'The Worlds I See', a book the Financial Times called 'a powerful plea for keeping humanity at the center of our latest technological transformation'. To her credit, she did not ritualistically immolate any symbols of non-anthropocentric technologies. But taken together with Sutskever's odd ritual, these two events are notable milestones in the wider reaction to a technology that is upsetting our self-image.

One of the other things I did in the last few months, I wrote and published a piece in Noema, that started off as a sort of side joke. What I did here is I took Elizabeth Kubler-Ross', five stages of grief—denial, anger, bargaining, depression, and acceptance—and used this as an interpretive typology for AI discourse. AI denial, AI anger, AI bargaining, AI depression, and AI acceptance. And you can read this at your convenience. The key idea here is that—in ways that are actually quite precious—AI as an existential technology is one that is already bringing about what Freud called Copernican traumas: that scientific ideas, and Freud was talking about Darwin, that decenter or destabilize our self-image as a species, as a collective subject, are ones that this destabilization is not just difficult, it's genuinely traumatic. And so I see the symptoms of grief not necessarily as bad news. I think it's a symptom that AI is a Copernican technology, and that the things that we will learn from it about ourselves are ones that would be precious indeed.

The simple way of thinking about this is that we're familiar with the idea that if you get more alignment, this will be a net positive outcome. We're familiar with the idea that if you have less alignment this will be a net negative outcome. Where the part that we're interested in, again, this is a starting point, not a conclusion, is what if actually less alignment is the path to a net positive? Also, just to extend the thought, we're also committed to the idea that alignment overfitting itself is an existential risk. That codifying and freezing a kind of uniform idea of human values or even a very diverse idea of human values, and then the strong presumption is that the extrusion of these human values and desires into a system that will accommodate them and reward them is, in the long term, the best way to think about the evolution of machine intelligence, is deeply suspicious.

In the short term, yeah, it's fine, you want to ask the AI to do something, you want it to do that thing and not something else. But I'm not really talking about that. I'm talking about the very long term idea that the more anthropomorphic an AI is, the more humanlike that it is, the better its evolution will eventually be. Or even worse, the appearance of anthropomorphism, the appearance of human likeness is even a problem. But there's also a deeper question about human values in and of themselves. I mean, have you met humans? The idea that you would want to build that planetary exoskeleton as something that thinks like we think, is, you know... I think you get the point.

What we think about here is what we call bidirectional alignment, which is the idea that, yes, in many respects, the evolution of AI must correspond to an align with, in a symbiotic way, the further development of human society into that long term future, but also the human society inevitably is going to continue to evolve in relationship to the appearance and artificialization of AI. And that transformation, that change of society towards the AI is sort of the opposite of alignment. We'll call this a bidirectional alignment.

Speaking of trauma, another way of getting to this is this: you're all familiar with Mori's idea of the uncanny valley, right? The idea that you see something that looks kind of human, but it's not quite human, and because it's not quite human, it kind of creeps you out. Well, if we're thinking about the way in which AI sees us, and we're kind of creeped out not by looking at something that's something else, but we're kind of creeped out by looking at ourselves through the lens of the other. We call this the inverse uncanny valley, where you're creeped out by looking at yourself through the eyes of the other.

This is the point. What we might think of as AI overhang, the idea that AI is capable of doing lots of things that our society and culture hasn't quite figured out how to absorb yet, happens not just in a person by person level or industry by industry level. It's a civilizational dynamic, that the appearance of machine intelligence is something that human civilizations, plural, haven't figured out how to engage with and how to absorb. But if we think of AI as an existential technology, not just an instrumental one, the protection of that overhang, the protection of the Copernican trauma, is an equally important part of anything that would dare to call itself AI ethics.

PART SIX: HAIID I.E. HUMAN-A.I.-INTERACTION-DESIGN

One of the areas when we're talking about is interface or interaction between human societies and artificial intelligence, the term we have for this more broadly is called human-AI interaction design or the acronym HAIID.

And some of the things we're trying to think through is how do you get past the bot paradigm? Text field is the only way of thinking about this. Are there other ways of metaphors and dynamics of interface reality that might be at work here, that might be useful? To set some of the ground rules: for us, interface isn't just GUIs. Interface is that which governs the condition exchange between two complex systems, and to add to this, also governs the condition of exchange between those systems. So things like GUIs are interfaces, but lots of other things are interfaces, though we may not recognize them as such. I think the important question for both philosophy and design, is what is the full range of interfaces between human intelligence and machine intelligence that is not only conceivable, but possible. And part of the reason this is so important is that that's where the rubber meets the road. That anything we might call bidirectional alignment is something that happens one interaction at a time, and builds up in patterns at a time. And so structuring the interface between these two modes of intelligence is how you're structuring and giving shape to the symbiosis. And that can't just be bots.

One of the other things we've been looking at in human interaction design is that in the contact between two kinds of intelligence and even those intelligences that are very similar to each other, is a kind of reflexive modeling. That one mind is not only thinking about the other mind, it's thinking about what the other mind is thinking about it. And that other mind is thinking about how the first mind is thinking about it. And in this mutual mind modeling you set the conditions for not only the possibility of exchange, but also the evolutionary dynamic that accelerates the appearance of intelligence in the first place, the basis of social intelligence. And so part of the anticipation of this is that AI evolution in the present also works through reflexive modeling at lots of different scales and in lots of different kinds of ways. And in that the emergence of social intelligence is, in fact, how this will work.

We're cognizant of the idea that a lot of the models in our minds that we may have of what's going on in the machine and what the machines as models of us may be completely fantasies. We might call these folk ontologies, but in many respects, it's all folk ontologies, it's folk ontologies all the way down. The idea that I have of what AI is or you have of what AI is, however well-informed it might be, is still a metaphor in many ways. It's still a kind of abstraction. But it's a useful abstraction, a heuristic abstraction. There's nothing wrong with the fact that it actually operates in this way. But still, I think we should be cognizant of that.

Okay, so in different ways, whenever you have an encounter between two complex forms of mind that are modeling each other, they're not always modeling each other in a 50/50 symmetrical ratio. Sometimes it's really very asymmetric. You may have a great deal of modeling, like a very intelligent creature might be modeling a less intelligent creature, it might be doing 70% of the modeling, while the other one's doing 30% of the modeling in terms of the complexity. And part of the idea is that it's the asymmetries in the social mind modeling that actually is generative of all of the proliferation of diversity and interesting things going on. But also the ratios of mind is interesting. It's not always one to one.

I showed a picture from Spike Jones' Her earlier and I assume everyone's seen this. My favorite scene in the film is probably the same as yours is the scene towards the end where Theodore is talking to Samantha, and he begins to realize what's really going on here, that his folk ontology of this agent is not what "she" is. And he asked her, 'how many people are you talking to right now?' And she says, '8316' and he realizes at this moment that what's been going on is very different to what he imagines has been going on. But here, this ratio of one to 8316, this is what we mean by the outnumbered ratios within this as well. This produces very different kinds of dynamics. We might say that in 2001: A Space Odyssey, Dave and Hal are in a one to one relationship. That's a one to one relationship. If you're a gamer and you're playing an advanced video game that has lots of AI-generated NPCs, that might be the other way around, that might be more like 8316 to one. They're kind of the inverse of the Samantha scenario. But the point is that it's the quality of the asymmetry of the social mind modeling but also literally the quantity of minds that are interacting through this interface that structures a lot of things that are possible, which inevitably gets you to the question of twinning and doubling and simulation.

And we're familiar with, in the present, also personas and fake personas. They're built on particular kinds of people. One of the ways normally thought is: I've got a digital agent of myself, of which there's a unidirectional causality, it's better if it's most like me. But inevitably, if you have a really complex agent like that, you also have a kind of parasocial relationship with it, and that the kinds of decisions that it's made come to recursively, reflexively impact not only the decisions that you make, but the ways in which you're thinking about that whole decision process. And it ends up being a kind of bidirectional recursive relationship, just like any other complex social ones. And furthermore, as those get more complex, like Samantha, they don't only talk to people—in any way in which you get the development of a complex ecology of digital twins that society emerges. And in many respects, the interesting conversations are happening between them, not between people that they are essentially avatars of.

One of the projects in the charrettes we did in the studio was based on a project we did in the first year's studio called Vivarium. And this is a project that deals with what are called toy worlds, which are simulated spaces in which AIs are trained to

do physical things that might be too complicated to learn in the real world, like driverless car or something like this, but also simple things where you're trying to figure out what would be the dynamic of the engagement of lots of these little agents in this space working together. You can simulate this and then kind of slowly move them out into the real world, the sim to real dynamic, which is sometimes very psychedelic for the AI, apparently. But also, you know, very lossy.

The part of what we're interested in here is also then how it is that many humans out here, AIs in here in lots of different ratios: one to many, one to one, many to one, and many to many, let's say, may involve a whole different kinds of dynamics of relationship between physical and virtual spaces, which is obviously really important for anything kind of cognitive infrastructure scenarios that have any kind of traction. Now, on the question of modeling, mind modeling, how do we think about what the AI is, what's called embedding visualization, is a pretty good, pretty reasonable kind of empirical, closer to ground truth way of thinking about, like, what the model is. And so what you see in embedding visualisation is the structure of the vector space in which words are associated with each other in different kinds of dynamics. So anything can be a token, so it may not 100% be words. But the idea is you're kind of seeing almost like a brain scan of the way in which the model is representing language.

And so one way to think about this, and this is a technical problem, because as you see here, embedding visualizations are kind of 1.0, in terms of the kinds of things that might be possible. People are doing brain scans as another kind of methodology, it's very interesting. Doing different kinds of editability, it's called, we're going and fixing weights on a sort of granular basis. But for us, the kind of interesting philosophical point of this is that what we're seeing here is our attempt at a representation of AI's representation of our language, which is our means by which we represent reality in the first place. So it's a representation of the representation of representation itself. A kind of triple mirror, in which we're looking at something, looking back at us, in which we are looking at the world like, whatever, how it sort of goes. But this kind of recursion and this kind of modeling is exactly where the real issues are going to be, very interesting things are going to appear.

And if this isn't clear already, we're not exactly following the straight and narrow in terms of thinking about where all this is heading and what the interactions between humans and AI might be.

And the other charrette we did at the beginning of the week was based around neurodiversity and human AI interactions, in which we try to think about in neither utopian nor dystopian sort of way. Lots of different ways in which what might seem like neurotic relationships between humans and AI actually might contain a kernel of something really generative and interesting that might turn out to be not so neurotic after all. And we generated numerous scenarios around this that are quite lovely and I'm happy to share them with you.

PART SEVEN: GENERATIVE/SYNTHETIC A.I.

55:39

Part Seven: Generative/Synthetic A.I.

All right, two more. Generative and synthetic AI. Generative AI. How do we make stuff with AI?

So this is generative AI circa 2015. The Deep Dream Project by Alex Mordvintsev, who's on Blaise's team, in which you have this kind of mushroom hallucination of the AI seeing dog faces in everything, because that's what it's looking for. And there's a longer explanation of this as well. But this is quite wonderful. I really miss this era of AI aesthetics quite honestly. But less than ten years later, generative AI also led to an actors strike and a writers strike in Los Angeles, just a few years ago. An entire culture industry was probably legitimately concerned that generative AI was going to transform the entire industry in its image. So from Trippy Squirrel to no more Hollywood in less than a decade.

A few months ago I did a lecture tour and research trip to China and one of the places I went was Kuaishou, which is a competitor of Douyin, or like TikToks. This is the text to video model that they're rolling out now. It's called Kling. It's like Sora, which you're familiar with, but it's not quite as good, but it's still pretty good. And their research group claimed to me that they anticipate by this time next year, as much as half of the content on their platform will be text to video generated. And so this arc of growth is not slowing down anytime soon. Here is a bit of words to live by: Always remember that everything you do from the moment you wake up to the moment you fall asleep is training data for the future's model of today.

Some of the things I've been working on thinking through a generative AI come from a piece I wrote for the catalog for Holly Herndon's show at the Serpentine coming up, and I'll share some of these things with you. One of the things I'm really interested in, the kind of politics of generative AI is what we might call the politics of derepresentation.

From the late 20th century to the early 21st century, the politics of representation was quite pervasive, and part of it was based on the idea of that by changing the representation of the thing is how you actually change the status of the thing, and

therefore the inclusive amplification of representation of many different things was a way of empowering those things by presenting them. And so the politics was sort of maximizing visibility.

In an era of data poisoning, data privacy, wherein entire countries are saying, for example, when Italy told OpenAI that they were going to sue them because GPT-4 had Italian data in it, the politics is not about ensuring agency and sovereignty by ensuring representation, but rather about a willful self derepresentation of minimizing visibility, of extracting yourself from the sample model of society as a way of controlling or attempting to control how that model might be used. And this is a very different kind of approach. And its long term implications may be quite strange.

One way of thinking about this is what we've been calling the Potemkin politics of the model, which is if to the extent to which large models that are pervasive in cities, governing those cities as forms of cognitive infrastructure, are doing the work of governance in the like 'small g' cybernetic sense of governance, that is the model is making doors open and close and things moving around, like just thinking kind of the low level logistical mechanics of an artificial environment as the form of governance. If you want to change the way in which that governance is happening, what you want to do is to change the way the model is representing reality. That the model has a model of the real, and it's governing the real through the model of the real. And so to change its governance, you change its model. Now, to change this model, you could be making that model better and making that model higher resolution, making that model more realistic. But another way of changing it is essentially to produce an artificial reality for the model—that it thinks what you're presenting to it is the real and transforms its representation of the real accordingly. So the production of something like the Potemkin Village version of the city, a performative act that is signifying itself for the sole purpose of being data and being data points that the model will incorporate as a way of rethinking the real is what we call the Potemkin politics of the model. And in many respects, I think this is where things are going.

A short story I wrote for our book on Hemispherical Stacks was called "The Myth of Blurope", and it was thinking about the longer term implications of things like the EU and Italy's retraction of its data. The idea that maintaining sovereignty would refuse to be represented, that in the long term the story takes place 20 years in the future, that there's some debate as to whether Europe even ever existed at all, because it doesn't show up in any of the data sets. There's this long process of self disappearance. It's a story, so relax.

You may remember the dustup a few months ago, with Scarlett Johansson's voice apparently being used for the new version of GPT, because Sam Altman thought it would be cool to have OpenAI's model talk like Samantha from Her. But it turned out things were not quite as they seem. Apparently it was not Johansson's voice, but another voice actor. It was recorded before Altman had asked Johansson to use her voice, and that Sky voice has been available for months before any of this became a kerfuffle. And so the story was not quite right. The story was wrong. As many AI critics told us at the time, the first version of this tale proved that in essence, we are all Scarlett. The platforms are stealing our very beings to sell it back to us, and that we must retain and reclaim our stolen dignity from the math monsters. But what then, as an alternative, does this new version of the story actually prove? Perhaps it proves that human culture is actually not made up of individually owned bits of property and speech acts, but is and always has been a manifold, a river from which individuality briefly emerges.

LLMs model the manifold, not the sign or the person doing the signifying, and this is understandably deeply confounding for many smart people, and for good reason. And it will for sure require a serious rethink of what the fuzzy continuum of human cultural intelligence really is. And the implications, I think, are particularly profound for reasons that should be obvious for generative AI and the ways in which it is trained on and productive of whatever culture actually is.

Another thing that we've been looking at in terms of what culture will be is in relationship to multimodal interfaces. Contemporary large language models are multimodal. You can take lots of different kinds of media: speech, text, audio, robot instructions, whatever, kitchen sink, and feed them into the model, and you get lots of those different kinds of things on the output—all those same things, plus even a longer list, and they can be recombined in different sorts of matrix. You put in a song, you get out a picture, you put in a novel, you get out a robot instruction, whatever. This is also, I think, very interesting in thinking about the kind of cultures that were built. It's obviously an interface problem because if you look at the apps on your laptop, they're all made for one media type at a time. Your video editing software is for video, your word processor for word processors. But if the software can basically do all of them in different combinations at once, how do we think through what we're making could potentially change very quickly in interesting ways. I think that the real problem, the real issue to think about with generative AI is that it's automating the process of making digital media is kind of missing the point. It's a bit like saying that TV automates radio. It's not just—generative AI is not digitized automating digital media, it's a framework by which you're in a completely different kind of relationship between image, text, sound instruction, execution is likely to emerge, but can only emerge if we have the interfaces to do it, one of which may be language.

You probably remember Jensen Huang from Nvidia, maybe a month or two ago, saying that classical programming is basically over, that English is the new programming language, and all the CS majors are, oh well. One of the ways in which this might be at least partially true is the way in which language encodes intentionality. And we're all familiar with this—I'm riffing from a piece by Boris Groys on this—we're all familiar with the idea of the death of the author, right? That the intentionality of the author is not the thing to look at in how it was constructed, but intertextuality. And I think that's also true.

The part of the way prompting works is by trying to encode intentionality and negotiate intentionality, so the thing knows what to make. And so the reason that language has this universal applicability is because it's a way in which we know how to encode intentionality, but there's lots of ways potentially to do that. We're also interested not just in the different modalities of prompting, but what different ways in which the training prompting dynamic, which we have right now, is very likely to be something quite different in the future, not only because of real time learning and infinite context windows, but also because there may be very different ways in which models come in contact with human intelligence.

That is, training is a way in which models come in contact with collective intelligence. It models the manifold. And then prompting is the moment it comes in contact with individual intention. Fine. But that can also potentially work the other way around, where training is something that's happening at a very granular and individual level. And increasingly it is, through real time learning and prompting. However, could also be to the extent to which it's the broader idea of encoding of intentionality. That's also what politics is. That's also how the intention of a society to how it wants to use this infrastructure to reorganize itself is manifested. And so that dynamic—I'm being really rough about this—the dynamic between something like encoding of intentionality through voting and encoding, intentionality through prompting may not seem so different, ultimately, as things come along.

One of the scenarios we've been thinking about is does that larger manifestation of generative AI lead towards convergence or divergence? Does it lead towards a world in which there is one song, that all the variety of everything on Spotify is kind of moving in this convergence towards the one song that everybody eventually listens to, and it just becomes the convergence song, a bit like the hamburger became the convergence sandwich or something. Or does it go the other direction where because there's such granular and detailed responsive interaction with every individual listener that it starts producing songs that are increasingly unique to your interests and aesthetics and intelligence, and tastes, that if there's lots of AI's in Spotify now, that basically what it will do is to create something like an infinite song that is just for you, that is less like a bunch of tracks in a playlist that are sequential things, but more like a kind of continuous, never ending song that's more like a river that you step into or out of at different times of a day. And perhaps it's so unique to you that basically no one else could stand to listen to it.

PART EIGHT: LANGUAGE ONTOLOGY

1:09:06

Part Eight: Language Ontology

Last point: language. We're talking about large language models, obviously.

One of the interesting things I think about this moment with large language models is that it's driving linguistics crazy. The number of friends I have in linguistics who are adamant and in many cases agitated that this is not language. That what's going on here, what it's producing and what it's doing is not linguistic thought. It's something else, and you're all barking up the wrong tree. And maybe there's a bit of defensiveness, but time will tell.

Let me come back to asking that question a little bit differently. You remember the old motto; to organize all of the world's information and to make it useful. Well, that means something a bit different than it did when this was first put out there. All the world's information is a lot of things, and it's not just the stuff that you might find on the internet. It may be a very interesting project, but it's one that's orders of magnitude more complicated and larger than it might seem.

A few months ago, many of you were probably there, I went to NeurIPS in New Orleans, and there were no less than 15 papers, and that's just the ones that I saw, concerned that we're running out of data. No more data to train our language models. But what they really meant was that there's no more easily and freely available English text to train things on.

Now, obviously, that's a tiny little slice of even what language is, let alone what the world's information is. And so what we really mean about running out of language in terms of information starts to open up some really interesting questions to think about what the relationship between language and data even is. And where do we distinguish? In other words, I don't think it's a controversial point, but the instantiation of large language models is going to change the fundamental ontology of what we even understand language to be, and it very well may be that the word language is the wrong word for what language actually is. That what language actually is is not what the word language in language thinks language is. I'm almost certain of it. 80%. Which also, as you know, opens the question of—is AI being made of the right

stuff? Is training AI predominantly on freely available English actually the right first step? The right initial condition for this? Why is it being trained on what humans do? Why is it being trained in what humans who speak English do? Why is it being trained on the text that humans who speak English produce? Why is it being trained on only freely available stuff that people leave out in the open? Is that the right initial condition? Are we building AI out of the wrong stuff and therefore producing the wrong data, and therefore limiting the potential of what AI actually is?

A line in one of my books—I was a bit critical of the Take Back Our Data movement in a very specific way—and I said, okay, great, here you go. Here's all of Facebook. There it is, all of it. You get all the vacation photos and cat videos and angry posts from crazy uncles. All of it. Now go solve climate change. It's the wrong data. It's the wrong data for that. And I suspect in many respects, and I don't think this is really a disagreeable point, that the space of potential data that we really need to be training the world's model of the world on needs to be much different in quantity and quality. One of the ways in which I think the ontology of data is changing is that language is now that which can be tokenized. Anything that can be tokenized can count as language. Maybe, maybe. That's one way of thinking about it, that this functionalization of language is not only changing what you do with language, it's changing what language is or the way in which we can expand the the word to mean something different, which means that that relationship between what is the boundary between what is language and what is data, what is language and what is calculation gets a little bit blurry, but it's always been blurry.

This is Sumerian cuneiform, of which you have a very nice collection at your museum. Which, of course, are the earliest forms of writing as we know of, are what? They're mostly receipts. They're accounting. The beginning of writing is accounting. All writing the humans have ever done, the entire canon, all canons combined. The greatest things you've ever read is basically a kind of fancy accounting, derivative accounting, which I love to remind my friends in the literature department. But what it means as well is that writing is a kind of calculation, at least in an original sense, or the dynamic between writing and calculation is not some newfangled perversion. It's actually fundamental to the relationship between language, the qualitative and quantitative, and that this is something I think we can reconsider and ask the question: is language more a form of calculation than calculation is a form of language? It's both, obviously, but that's my point.

This opens up a lot of different questions and we don't have the answers to these, but these are some of the ways we're thinking about that. And that is we can think about languages, it's like some of the most interesting research on large language models, and there's a number of people working on this, is how it is when you model the language and you zoom out, you come to the conclusion that languages are, in essence, topographically similar to each other.

There's a kind of isomorphic structure to human languages. And so, you know this, Chomsky's old line that if aliens were to come to Earth, that they would think humans all speak the same language is not entirely wrong. Even though Chomsky is obviously wrong about other things about AI. You've got English and Spanish and German, and Swahili and Finnish and Swedish and Japanese and Chinese, and all kinds of things, that roughly take the same kind of shape.

The question that maybe to ask is, is that bounded shape that they already take actually the limit of language itself? And is the bounding of that shape that they already take, the shape that it is because somehow language actually is a good approximate representation of reality. And so reality has selected for and constrained the evolution of language to this particular shape? Possible. Or is there a much bigger space of alignment between language and data, language and information, of which the particular structures of language that we're familiar with, that we have been speaking and thinking through is actually just a tiny little sliver of the larger space of what language is, which I have to be inclusive, ultimately, of information that the boundary between information and language is actually quite, quite porous. Maybe. Or maybe there's somewhere in between that language is larger than the languages that we have, that would include other kinds of non-human languages.

We had a guest talk from Gašper Beguš the linguist who works on CETI—the project that's using machine learning to analyze whale songs and whale language. Maybe language is bigger than our version, but it's still a subset of the larger space of information. That language is essentially a subset of bio semiotics, more broadly. Could be, we don't know. But nobody knows. But these are, I think, some of the questions that we're kicking around at the very least.

But the idea then is that if AI is the way in which the planet is making a model of itself, then is that model really bound by language as we know it, or can the space be much larger than it can be? Looking at this, of course, through the lens of what language is itself. Now, all of this is evolving not from the past, in the present, but also into the future. And so as I showed you in that chart that AI is not the end point, but rather the thing that is a scaffold for whatever comes next, we might ask this question: what is language? And the artificialization of language is a scaffold for what? Not only what is it, but what may be the answer to the question of what it is, is the same as the answer to the question of what it's a scaffold for?

And it may be quite weird, obviously. Several years ago we did a visit to the Sony AI center, and all of this has, of course, gotten much more interesting since then. But one of the things they showed us was a language that they developed for these industrial robots. What you're seeing here in that strange, finished looking language is one robot telling the other robot to pick up that yellow pylon and put it on top of the box over there. But you know where this is going. What happened within 72 hours was that that language evolved so quickly between the robots that they were able to do things and coordinate, but the researchers had no idea what they were saying. This I think quite obviously is where things are beginning to head, that the future of language is one that's quite alien to us, almost inevitably. That language always takes on a life of its own. And that's good.

We could imagine this scenario where right now you've got roughly 8 billion humans speaking human languages of various difference and similarity. But that very soon most of the speakers of human languages won't be human. Most of the speakers of human languages won't be human. In number and quality. We, the 8 billion people, will be simultaneous with 80 billion, give or take, non-humans speaking human languages. And what they do with those human languages will be rather different than what we do with the human languages. But the dynamic between the two isn't just of separation. So if that question is, okay, how are all the languages similar or different to themselves? And how can large language models allow us to analyze language, the representation of the representation of representation itself? That's, we can say, meta linguistics. Xeno linguistics would be the language that all the ways in which the non-humans are speaking human language and what might call allolinguistics would be the impact that that has on the way we think and speak.

The ways in which the AI language will evolve through AI will kind of reflexively and recursively change how we think and speak. That's where we are, that's the scenario we're playing with.

PART NINE: CONCLUSION: POST-GRIEF

1:20:15

Part Nine: Conclusion: Post-grief

Alright. I said we were giving some ideas that we're playing around with and working through. And that's true.

But before we end I owe you at least some provisional conclusions, which after my five stages of AI grief piece we might call post grief. And so, depending on your tastes, let me give you a sad conclusion and a happy conclusion. And then you can pick which one you prefer.

First, the sad conclusion.

The now essential general technology of artificialization is computation. The way in which humans artificialize the world is through computation. Vision, biology, intelligence, environments. Computation is the general medium of artificialization. As computation is now a central technology for artificialization, it is becoming artificialization itself. It is artificializing the locus of artificialization itself. Re-scaffolding and re-technologizing the structure and dynamic of intelligence in its own right. However, in the shadow and context of what people used to call the Anthropocene, we might ask if complex intelligence is ultimately adaptive or self extinguishing in the long term? Obviously, evolution selects for artificializing intelligence so well, really well in the short term, that it may burn itself out. That flywheel effect of self amplifying intelligence capacity for artificialization is, that's the dynamic of the big acceleration of time and complexity. But in fact, what we see is the Anthropocene, maybe a limit condition for it. So, maybe not. Well, if we were to demonstrate that it's not, and find our way out of that potential loop, we would have to answer a few questions. What would make advanced intelligent artificialization adaptive in the long term, symbiotic or symbiotic with planet? What are the preconditions for that adaptiveness? How can these preconditions be artificially realized? They're not just going to evolve necessarily. And the question is that the real Copernican trauma at hand is to try to solve that question of whether the short term and long term adaptiveness of artificializing intelligence actually have totally different fitness functions. So that's the sad conclusion.

Let me give you the slightly happier one.

This is also from the 'Stages of AI Grief' piece by the way. When James Lovelock wrote his last book 'Novacene: The Coming Age of Hyperintelligence' he knew that he was dying. And in the book, he concludes his own personal life's work with a chapter that probably startled some of the more mystically minded admirers of Gaia Theory. In the end, he calmly reported that Earth life as we know it may be giving way to abiotic forms of life and intelligence. And as far as he is concerned, that's just fine. He tells us quite directly that he is happy to sign off from this mortal coil, knowing that the era of the human substrate for complex intelligence is giving way and making for something else. Not as transcendence, not as magic, not as leveling up, but as simply a phase shift in the very same ongoing process of selection, complexification and aggregation that is life, that is us. There are then, non-grief ways of thinking through a philosophy of artificialized intelligence that are neither optimistic nor pessimistic, utopian or dystopian. Part of what made Lovelock at peace with this conclusion, I think, is that whatever the AI Copernican trauma turns out to mean, it

does not mean that humans are irrelevant, or that they are replaceable, or that they are at war with their own creations.

Advanced machine intelligence does not suggest our extinction, neither as noble abdication nor as bugs screaming into the void. It does mean, however, that human intelligence is not what human intelligence thought it was all this time. It is both something we possess, but which also possesses us even more. It exists not as individual brains, but even more so in the durable structures of communication between them.

For example, in the form of language. Like life, intelligence is modular, flexible, scaler extending to the indigenous work of subcellular living machines and through the depths of evolutionary time. It also extends to much larger aggregations, of which each of us is a part, and also an instance. There's no reason to believe that the story should or will end with us. Eschatology is useless.

The evolution of intelligence does not peak with one terraforming species of nomadic primates. For me, this is the happiest news of all. Like Lovelock, grief is not what I feel. Thank you.