

What does it mean to ask machine intelligence to “align” to human wishes and self-image? Is this a useful tactic for design, or a dubious metaphysics that obfuscates how intelligence as a whole might evolve? Given that AI and the philosophy of AI have evolved in a tight coupling, informing and delimiting one another, how should we rethink this framework in both theory and practice?

The emergence of machine intelligence must be steered toward planetary sapience in the service of viable long term futures. Instead of strong alignment with human values and superficial anthropocentrism, the steering of AI means treating these humanisms with nuanced suspicion, and recognizing its broader potential. At stake is not only what AI is, but what a society is, and what AI is for. What should align with what?

Synthetic intelligence refers to the wider field of artificially-composed intelligent systems that do and do not correspond to Humanism's traditions. These systems, however, can complement and combine with human cognition, intuition, creativity, abstraction and discovery. Inevitably, both are forever altered by such diverse amalgamations.

In *After Alignment*, Benjamin Bratton discusses shifts from AGI to artificial generic intelligence, the importance of recursive simulations, the decentering of personal data, the challenges of AI in science, intelligence as an evolutionary scaffold, the limitations of mainstream AI ethics, and why a planetary model of synthetic intelligence must drive its geopolitical project.

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

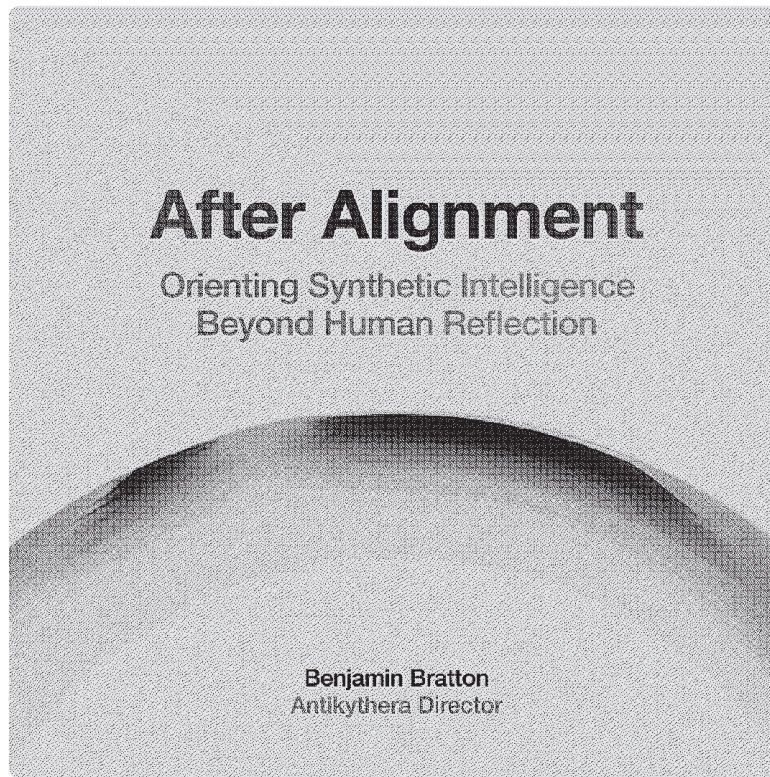
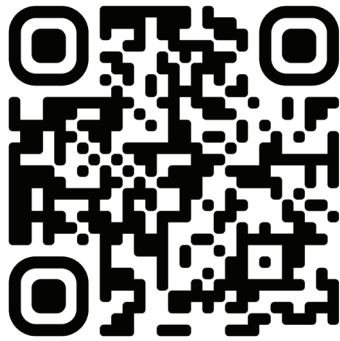
The film, *After Alignment*, delves into the complexities of aligning artificial intelligence (AI) with human values and intentions. It distinguishes between *alignment* as a practical tool for directing AI behavior and *Alignment* as a broader metaphysical concept, suggesting that the latter may be insufficient for guiding AI's long-term integration into human society.

Bratton critiques the tendency to idealize “human-like” qualities in AI, asserting that this constrains our understanding of existing forms of machine intelligence. Instead, the piece calls for developing a more nuanced vocabulary to analyze and speculate on the “weirdness” of AI, focusing on what it reveals and does, rather than merely adhering to precedent models.

The film posits that AI's unique insights—its *epistemic overhangs*—should not be viewed as anomalies but as opportunities for humans to gain new perspectives on themselves and their world. This two-way alignment encourages a symbiotic relationship where AI not only serves human purposes, but also One of Antikythera's projects, HAIID (Human-AI Interaction Design), catalogs various models and patterns of human-AI interaction. This compendium serves as a resource for understanding and generalizing the diverse ways humans and AI systems can co-evolve.

The talk emphasizes that traditional philosophical frameworks may be inadequate for addressing the challenges posed by advanced AI. Instead, it advocates for a speculative philosophy that emerges from direct engagement with computational technologies. This approach seeks to develop new concepts and vocabularies that can better capture the nuances of AI's role in society.

After Alignment calls for a reimagining of the human-AI relationship, one that moves beyond control and compliance towards mutual growth and understanding. It suggests that by embracing the unique capabilities of AI, humans can gain deeper insights into their own nature and the evolving landscape of intelligence.



The history of AI and the history of the Philosophy of AI are deeply intertwined, from Leibniz to Turing to Hubert Dreyfus to today. Thought experiments drive technologies, which in turn drive a shift in the understanding of what intelligence itself is and might become, and back and forth.

00:54

History and Philosophy of AI

But for that philosophy to find its way today, and for this phase of AI, it needs to, that finding its way needs to include expanding from the European philosophical tradition of what AI even is, and the connotation of this. From the connotation of artificial intelligence drawn from the Deng era in China was as a kind of relation to industrial mass mobilization. The Eastern European model includes what Stanislaw Lem called existential technologies, just as in the Soviet era it meant something more like governance rationalization. All of these contrast with the Western individualized and singular anthropomorphic models that dominate contemporary debates still today.

To ponder seriously the planetary pasts and futures of AI, we must extend and alter our notions of artificiality as such, intelligence as such, and must not only draw from this range of traditions, but also, to a certain extent, almost inevitably, leave them behind.

What Turing proposed in his famous test as a sufficient condition for intelligence, for example, has become instead solipsistic demands and misrecognition. To idealize what appears and performs as most “human” in AI, either as praise or as criticism, is to willfully constrain our understanding of what machine intelligence is as it is.

And this includes language itself. Large Language Models and their eerily convincing text prediction capabilities have been used to write novels and screenplays, to make images and movies, songs, voices, symphonies, and are even being used by biotech researchers to predict gene sequences for drug discovery. Here at least, the language of genetics really is a language. LLMs also form the basis of generalist models capable of mixing inputs and outputs from one modality to another, you know, interpreting what an image it sees so it can instruct the movement of a robot arm and so forth. Such foundational models may become a new kind of public utility around which industrial sectors organize what we call cognitive infrastructures.

So what about speculative philosophy then? Well, I honestly don’t think that society at present has the critical and conceptual terms to properly approach this reality head on. As a coauthor and I wrote recently, “reality overstepping the boundaries of comfortable vocabulary is the start, not the end, of the conversation. Instead of groundhog-day debates about whether machines have souls, or can think like people imagine themselves to think, the ongoing double-helix relationship between AI and the philosophy of AI needs to do less projection of its own maxims and instead construct newer, more nuanced vocabularies for analysis, critique, and composition based on the Weirdness right in front of us.”

And that is really the topic of my talk, the weirdness right in front of us and the clumsiness of our language to engage with it.

Toward that, let me say that, again, that the impasses over whether machine intelligence has mind or sentience or consciousness, are in fact impasses in our language and our imagination more than they are in what is actually happening, has happened

and will happen. The productive interest instead may have less to do with how AI adheres to precedent models than what it reveals and does about the limitations of those models.

So I say *reveals* and *does*. Why the split? Well, instead of presuming that ideas are first formed, and then tools are wielded to act upon them, we may observe instead that different tools make different ideas possible. It's not just that they invite different dispositions towards the world; they literally make the world conceivable in ways otherwise impossible.

Per Lem -Stanislaw Lem that is- and his notion of the epistemological and instrumental technology distinction, we might say that some kinds of technologies have the greatest social impact in what they do and enable in artificially transforming the world. These are instrumental. Others, however, have greater social impact in what they reveal about how the world works. These are epistemological technologies. Telescopes and microscopes are good examples.

05:27

Epistemological and Instrumental Technologies

Yes, they allow perception of the very large and very small, but more importantly they enable Copernican shifts in self-comprehension, grasping our very selves as part of planetary and indeed extraplanetary conditions. With such shifts, it was possible to orient not only where the planet is but thereby where and when and what "we" are, and of course also thereby putting into question that collective pronoun itself. Taken together, again, these may be called *epistemological technologies*.

06:33

Copernican Shifts

And it is certain that computation is artificially transforming the world, in the form of an accidental megastructure that shifts politics and economics and cultures in its own image. However, computation is also an epistemological technology that has and does and will reorient the course of what a viable planetary condition may be.

07:15

The Stack

In fact, we may say that the planetary as such is an image that emerges via computation, via, for example, climate science, which is based, of course, on planetary sensors and models and, most of all, supercomputing simulations of the planetary past, present, and future. In other words, the very idea of climate change is an epistemological accomplishment of planetary computation, and thus so indirectly is the notion of the anthropocene, and of humanity as a terraforming subject. And that is what is at stake.

07:40

Planetary Computation

So full disclosure then in this regard, my own approach for a Philosophy of Technology can then be understood as, in a sense, an inversion of the malaise carefully lamented by Heidegger, for whom technical reason's alienation of the intuitive "givenness" of the world is its and indeed our downfall. Whereas for me I think he has it backwards. That alienation, that Copernican weirdness, achieved through technological mediation of our cognition, has been and will be a path to access anything called Being: once again, where we are, when we are, where we are and how we are.

08:22

From Philosophy of Technology to Technology of Philosophy

But clearly AI is not only disclosing these, it is also forcing us to question them. We then experience different kinds of, if you like, AI overhangs and arguably underhangs.

An AI application overhang means that the technology is capable of doing things that a society has a hard time integrating, modeling, adopting for any number of good or bad reasons.

An AI application overhang means that the technology is capable of doing things that a society has a hard time integrating, modeling, adopting for any number of good or bad reasons. On the other hand, in the Antikythera program, one of the Metascience projects by Darren Zhu, Will Freudenheim, and Imran Sekalala calls an *AI epistemic overhang*, meaning that AI is capable of discovering things, knowing things, disclosing things that human science has a hard time modeling and integrating and adopting. The latter, we would argue, is not just an issue for science. In its generality, it is in many ways the focus of this talk.

So first about Alignment.

PART ONE: ABOUT ALIGNMENT

What does it mean to ask machine intelligence to "align" to human wishes and self-image? Is this a useful tactic for design, or a dubious metaphysics that obfuscates how intelligence as a whole might evolve? Given that AI and, as said, the philosophy of AI have evolved in a tight coupling, informing and delimiting one another, how should we rethink this framework in both theory and practice?

10:31

Alignment

Or let me put it somewhat differently. If The Stack describes the topology of planetary scale computation, the question it implies is, what is planetary scale computation for? We might insist that the emergence of machine intelligence must be steered toward a kind of planetary sapience in the service of viable long term futures. And for that, instead of strong alignment with human values and superficial anthropocentrism, the steerage of AI means treating these humanisms with some nuanced suspicion and recognizing instead a broader potential. At stake is not only

11:00

Planetary Sapience

what AI is, but what a society is, and indeed what either one is for and what should align with what?

The term synthetic intelligence in our parlance refers to the wider field of artificial-composed intelligent systems that both do and do not correspond to Humanism's traditions. These systems, however, can of course complement and combine with human cognition and intuition and creativity and abstraction and discovery. But as such, both are forever altered by such amalgamations.

11:58

Forever Altered

Now, machine intelligence itself may or may not be strictly speaking "artificial". If we mean artificial as something that is composed deliberately by some kind of precedent intelligence, then AI is a form of machine intelligence that is so composed. But it is perhaps not so simple. We recognize every day that there are forms of machine intelligence that are genuine and yet "evolved" without deliberate design; just look around the city. And second, we can zoom out and see that the intelligence that does any artificializing is itself evolved, and so its artifacts are then also part of this evolutionary phylogeny.

12:31

Machine Intelligence
May or May Not Be "Artificial"

The primary significance of this for the present talk is that we should not, repeat *not*, see "AI" simply as a direct reflection of human ideas, culture and economics, nor, vis-a-vis alignment, should it be.

13:23

Reflectionism

Put directly, the extent that it directly reflects human culture is not a goal nor is it a reality. The extent to which it departs from human culture is not what we might name a disaster nor is it a hypothetical. That departure is in fact our reality.

That human intelligence must or should orient -which is our preferred term- AI toward viable planetary futures is essential, but again, that viability does not arrive simply from making AI resemble us, or admire us, or be subservient to our wishes. To the contrary.

14:02

To Resemble or Admire

Now, there is obviously a Venn diagram overlap between AI Ethics on the one hand and AI Alignment on the other. But, there is also a kind of an impasse between them, or at least in principle, a contradiction between their visions. AI Ethics, particularly in guises associated with popular pundits and so forth, insists, when coaxed, that "AI" is simply the reflection of human societies, from unjust biases and unequal economic systems that produce it. It seeks to demystify AI as "just us" : nothing more, nothing less. As a "theory" it asks us to identify, in principle, any technology, and especially AI, as ontologically artefactual.

14:31

The Contradiction

AI Alignment, on the other hand, and again my abstraction of these correlates more to the popular and populist guise than any deeper serious research, Alignment may hold that the potential existential or at least serious risk of AI is based on the fact that it is now so deeply divergent from human cultural values and norms. And that securing a safe future means actively gluing it to those values and norms.

15:27

Shadow Puppets

So you see the theoretical problem. How can AI be both automatically reflective of human biases and values and dangerously unreflective of human biases and values? How can the positions that straddle my cartoon binary hold the apparently contradictory conclusion at once? How can we observe that, again, that AI is us, and this is bad, and simultaneously that it is not us, and this is also bad.

Well, it can be both, but only if we would radically qualify and specify what we mean by "alignment" and allow for alignment such that AI not only bends to social norms but also for which society evolves, in a positive sense, in relation to the epistemological and instrumental affordances of AI.

So before I go into a little bit more detail about what I envision, let me further contrast and clarify what I don't envision, what I don't mean.

The possibly very sensible perspective that AI and potential AGI pose an existential risk and so therefore should be the focus of planetary concern for geopolitical and geosocietal debate, has not always been as well represented in the public sphere as it should be.

AI moral panic overwhelms imaginative reason in what amounts to several simultaneous AI moral panics competing for attention, oxygen and hegemony.

17:08

Moral Panic

These range from rather predictable American Culture War templates that focus less on what is said than who is saying it, to a strange inversion where some of those most well known for rapturous millenarian visions of AI have rotated to apocalyptic eschatological ideas without missing a beat. From The Singularity to The Unabomber and back is the new horseshoe theory of AI politics: the hype / doom binary imploding into itself.

17:43

AI Horseshoe Theory

Elsewhere, but not too far away, many public intellectuals spent much of the Spring of 2023 in a perhaps well-meaning piety game to see who could say that we are in fact "more fucked". "You say we are fucked because of AI, well I say we are more fucked than thou." Sometimes the game eventuated in public letters of concern of varying quality and intellectual legitimacy, I think, but almost all with signatories that included very good and smart people.

18:18

More Fucked Than Thou

The most dubious of these, however, called for “Those In Charge” to push the imaginary Red Pause Button on AI “until we can figure out what’s going on” as the man said, knowing full well that no such button exists, that bad actors will not play along, that themselves perhaps stand to benefit down the line for having signaled their concern thusly. Most importantly, unlike the aliens in a sci-fi movie who land and say “take me to your leader”, when it comes to the serious risks identified there is no “Them” to petition. Who is in charge is the right question, and I would like to think that posting that question was the real point and hopefully impact of the letters.

18:55

“To Those In Charge”

That said, it’s clear that the present discourse around AI is, ironically enough, exemplary of the kind of discourse that the discourse around AI is warning us about: tribal, hyperbolic, truthy, unnecessarily dominated by whoever talks loudest and says the most outrageous thing, all amplified by advertising machines.

19:52

AI Discourse About Discourse of AI

PART ONE: AI DENIALISM AND AI ABOLITIONISM

Some critics may even go so far as to insist that AI is neither Artificial nor Intelligent. They say it is not artificial because it is made of physical materials, by people for specific purposes, which is the very definition of artificial. They say it is not intelligent, because it is merely modeling and solving problems, which in many significant ways is a good shorthand definition for a general theory of intelligence that is inclusive of but not exclusive to what it feels like to be human.

20:19

“It’s Not A or I”

Others may go a step further and advance a position that we might call AI Denialism. That is: AI doesn’t really exist. Don’t be fooled: it’s “just statistics”, “just gradient descent”. This is a bit like saying a symphony doesn’t exist, it’s just “sound waves”, or that food doesn’t exist, it’s “just molecules”. All of these are trivially true, but this AI Denialism is not remotely helpful in addressing the concerns it purports to stand for by making this particular case. The thing is, it is usually advanced as part of a political position in relation to the economics of AI, often a quite legitimate one, that then raises the stakes to an ontological claim. And so, it’s perhaps difficult to climb down from Denialism because it seems to put the validity of the politics in question.

20:59

AI Denialism

Elsewhere, AI Denialism dovetails with what we might call AI Abolitionism: AI does not really exist but should nevertheless be abolished.

Now, there is a lot to unpack here to do these positions justice, to systematically critique the critique of AI in this way would take at least a few other lectures, but let me then just so offer the punch line of what those lectures might be, because it’s also the punch line of this lecture.

Sociomorphism, that AI is or should be the reflection of human society, is the logical extrapolation of anthropomorphism, that AI is or should be the reflection of a single human. Both of these or neither of these is a real alternative to a California Ideology’s planetary hegemony; they are, in fact, the pinnacle of it.

22:38

Sociomorphism

The strong anthropomorphic view of AI goes back at least to the Turing Test, where what Turing offered as a *sufficient* condition for identifying machine intelligence became a *necessary* condition. That is, unless the AI could perform thinking the way that humans think that humans think, then it’s disqualified.

23:05

Turing Conditions

This idealization of what we could call Reflectionism, manifested as well in the psychologism of Human Centered Design -which proved a very mixed bag, as it turned out- and is present in the now contested terrain for Human Centered AI, Humanistic AI and so on, which are perhaps posed to make many of the same errors as HCD.

23:29

Reflectionism Again

I will talk about what we call Human AI Interaction Design, or *HAIID*, in a moment. At Antikythera we are very invested in this idea, not despite all its weirdness and complexity, but *because* of it. We are interested in the weirdness as well that is ensured by attempts to eradicate the badness. The last decade of AI ethics surely prevented a lot of horrible things. Not to mention how alignment researchers themselves contributed to many of the core technologies that we now make use of: from scaling laws to RLHF in particular.

23:59

From HCI to HAIID

And yet also, at the same time, “ethics” ended up dovetailing accidentally with corporate brand concerns to give us LLMs that are lobotomized to never speak about sex and violence in any meaningful way. We have prudish AI’s. We all foresee how bad it could get if the worst human preoccupations were directly “aligned” with the power of Foundation Models.

24:46

GPT No Sex Please

But at the same time, we also think about the critical role of sex and violence in the evolution of animal intelligence, including ours, and so recognize the weirdness of machine intelligence evolving with these topics as unspeakables.

25:13

Sex and Violence

Again, to be clear, my propositions on this are not intended to be posed at the expense of the research in AI ethics and Alignment, but rather actually in concert with their conclusions that Ethics and Alignment as such are together *necessary* but *insufficient* frameworks for the long-term orientation of machine intelligence. That, in other words, Alignment overfitting is real.

25:31

Necessary and Insufficient

My concern, however, is that exhaustion with tech solutionism gives way to self-congratulatory parades of political solutionism, that is now overflowing op-ed columns, trade books and yes, schools and universities. On the Continent, the inability to grasp how planetary computation upends 18th century forms of Westphalian citizenship leads to Regulatory solutionism: AI Laws that address yesterday's problems, the EU forever running to where to ball no longer is, trying to shoehorn planetary dynamics into citizen scale policies. For example, its permanent focus on individual citizen data as the core locus of concern and governance is, in a post-pandemic world, probably the wrong lens.

26:01

Political Solutionism

The way out of this is to cut the knot of the weakest forms of Reflectionism, which we might define as the moral and practical axiom that AI does or should be ontologically anthropomorphic. Not only are technologies not exhausted by the projection of social relations upon them; they are capable of forcing new practical concepts that contravene those social relations in the first place.

27:03

Concept Forcing

As such, the harm, to use the parlance of the day, is not only in what Reflectionism directs our attention to, but equally, perhaps more so, what it directs our attention from. AI represents an existential risk and existential potential in both senses of that term. By this I don't mean that it may or may not kill us all but that it may or may not disclose to humans and to animal intelligence -to planetary intelligence- fundamental truths about what we are, what their existential condition really is. This is the Copernican risk / reward calculus, one that is neither messianic nor apocalyptic.

27:38

The Harm

Now, you may hear my criticism, such as it is, of shallow AI anthropomorphism, and say, yes, but as it turns out AI does reflect human intelligence in some important ways. So, against a shallow anthropomorphism that insists that AI present itself as thinking how humans think that humans think, there is also a deep anthropomorphism, or better, a deep *biomorphism* that may correspond to how humans and other animals do really think, even if they don't experience thinking in that way. This is not only true, it's profound. And this is as I say, what shallow Reflectionism muddles.

28:30

From Anthropomorphism
to Isomorphism

For example, very recently, research in how AIs discern edges in images, for example, directed researchers to an unidentified, as-yet-unidentified, unspecified neuronal mechanism in human brains that perform more or less the same thing. As I will discuss in relation to one of Antikythera's projects, AIs are becoming a kind of experimental organism -like a lab rat- in which it is possible to test for human conditions and responses. This would not work if there was no fundamental correspondence. But what is and isn't the quality of that correspondence is of genuine philosophical and practical interest.

29:22

Model Superorganism

We also see at perhaps higher levels of abstraction that iterative predictive dynamics of transformer models do correspond with the iterative predictive dynamics of biological neurons. This was not the plan, but there it is. So yes, actually, you are a stochastic parrot, after all. Always have been. But you should not take that as the insult that the authors of that infamous paper perhaps intended it to be.

30:12

Yes, You are a Stochastic Parrot

Iterative stochastic prediction, thinking through the recursions of mental simulation and embedded in body perception is how humans made all the things that you hold most dear: literature, music, science, and so on. Parrots, by the way, are actually very smart and creative, so they are kind of a lousy token species for mindless repetition. But that's another thing. This deeper correspondence between iterative stochastic prediction and artificial / natural systems is technically an anthropomorphism, but as said, probably better to call it called biomorphism and suggest then a different AGI, an artificial generic intelligence.

30:45

AGI (Artificial Generic Intelligence)

Now, and this is really the point... Emphasis on the correspondence between AI and the manifest image of human thought, intelligence and culture comes at this terrible price: obscuring the real and profoundly significant correspondence between animal and machine intelligence that do not already register in common cultural norms but which could orient those norms to the underlying reality from which they emerge. A different bidirectional path of and for alignment.

31:36

Bidirectional Alignment

Notice I say from which they emerge, as opposed to the reality that emerges from those cultural norms. This is where perhaps there are some points of difference in our approach and some others on offer in the Humanities. It comes down to something rather fundamental of what is inside of what. The "planet makes worlds" or "worlds make planets".

32:17

Which is Inside of Which

I say that we must avoid what obscures the deeper and more philosophically challenging ways that AI does think like brains but simultaneously does not orient itself around Humanist norms, which thus distances human brains from human values in uncomfortable ways. Is this the real point of contention and unease?

33:16

"Nothing Outside the Text"

In the most extreme versions of Reflectionism, what is being defended, I sometimes wonder, seems like a kind of politico-theological conviction that there is nothing outside the text, as they used to say: nothing outside the sociological interpretation of technology, the political economy of science, the reality of culture as determinant of reality... Nothing causes culture but culture itself, culture causing culture which is caused by more culture, and thus anything, including AI, is intrinsically a reflection of that culture and nothing more. We might call this social reductionism and cultural determinism, which for all its lip service to posthumanism can be the most militant guise of humanism.

Now, obviously AI as it exists is fortunately and unfortunately a reflection of the cultures that produced it, but, and here is perhaps the critical fork in the road: it is not nor should it be only a reflection of culture. It is more. We are more. AI itself, and more importantly the qualities or reality that are certain to be revealed by AI, all the Copernican twists to come, are things to which culture must and will align, not only something that must and will align to culture.

34:17 To Which Cultures Align

So, two conclusions.

PART ONE: ALIGNMENT CONCLUSIONS

To sum: AI is an existential technology and as such must align in both directions: AI aligning to culture's wisdom, culture to AI's disclosures. So, specifically:

First, lower case alignment should be seen as a tactic for making machine intelligence's instrumentality, making it cohere to agency and intention, to make it work. But uppercase Alignment, and the attendant metaphysics, is an inadequate grounding for the long-term orientation of machine intelligence by animal intelligence.

35:34 Conclusion One

Second, a two-way alignment is possible and desirable. AI's epistemic overhangs, things it knows and implies for us to know that we have a difficult time grasping and accommodating and incorporating, are not pathologies: they are in fact the deeper point of AI. And so between AI as Generic Intelligence, AI as an experimental super-organism (per one of our Antikythera projects called The Ends of Science by studio researchers Darren Zhu, Will Freudenheim, and Imran Sekalala), AI becomes an uncannily productive sort of mirror. But it is not a mirror reflecting what we think we are because we can see it and feel it, but rather a mirror of what we are but cannot see and cannot feel, at least not yet.

36:01 Conclusion Two

PART TWO: ABOUT ANTIKYTHERA

Now. I would like to specify and ground this in some of the work that we've done to try to explore these ideas and many others in the Antikythera program. And to tell you a little more about the studio, but more importantly about the work.

37:08 The Program

All projects that I'll show you in a kind of summary coming up were all just completed at the end of last week where they were first privately shown in Los Angeles. And again, we will be back in London in the fall to do a bigger showcase around these.

Antikythera, as Stephanie signaled to you, is a research program, a think-tank of sorts, for the speculative philosophy of computation. It is supported and housed at the Berggruen Institute. We're pleased to be joined by Nils Gilman and Bing Song from the institute here tonight, Stephanie is the Associate Director, Nicolay, also here, my long-time friend and collaborator since Strelka, is our Studio and Design Director, and we are also held afloat by Case Miller and Emily Knapp, and growing, quite growing, every day it seems. The program includes 70-plus affiliated researchers from around the world and various universities, and has just completed in our last phase with 12 studio researchers who completed this studio cycle. It is, in essence, a program that seeks not only to map planetary computation, but to ask and provide some provisional answers to what planetary scale computation is for.

Now, as I have already said, philosophy -and more generally the project of developing viable concepts about how the world works and thus thinking about how the world works- has always developed in conjunction with what technology reveals and does and thus what is possible to think. And so at least in that regard, I am something of a technological determinist but only if we expand the definition of technology to its properly expansive scope.

38:48 What Technology Reveals

Here's the thing. At this moment, technology and particularly *planetary scale computation* has outpaced our theory. The response, as I have hinted tonight, is to some extent to force comfortable and settled ideas about ethics and scale and polity and meaning onto a situation that not only calls for a different framework, but is already generating that different framework.

39:19 Theory Outpaced

So instead of simply applying philosophy to the topic of computation, we start from the other direction and produce ideas -the speculative- from the direct encounter of making things.

That being said, the Antikythera program's real interest is not so much in calculation and formalization, quantification, or interoperability as such, than it is about how computation provides orientation, navigation, cosmology: in essence, planetarity.

The inspiration for the name comes from the Antikythera mechanism, first discovered in 1901 in a shipwreck off the Greek island of said name, and dated to 200 BC. It is, perhaps apocryphally, the first computer, but it is certainly a primordial computer.

40:24 The Mechanism

But it was not simply a calculator; it was also an astronomical machine, mapping and predicting the movements of stars and planets, marking annual events, and orienting a naval culture upon the surface of the globe.

So it not only calculated interlocking variables; it gave a comprehensible orientation of thought in relation to its astronomic predicament, enabling prescriptive thought to act in relation to this revealed circumstance.

So beyond forms of computation that are already perceivable in natural systems, artificial computation such as this is a kind of world ordering, a foundation for what would become complex culture. And that is really the core of it. For our initiative, the name Antikythera refers to computational technology that discloses and accelerates the planetary condition of intelligence.

So let me go a bit deeper into some of the themes and ideas of the program and show you some of the work which will be playing in the background as I explain a little bit where it came from and what it's up to.

PART TWO: HAIID

Not all of the projects, but several of them speak directly to questions of AI and indeed to questions of AI alignment rather directly. The project *HAIID* (by Antikythera Studio researchers William Morgan, Sarah Olimpia Scott, Daniel Barcay) is what you see here. It is an ever-growing catalog of existing and almost-existing modes, positions, and syndromes of HAIID that allows us to map and generalize the space. It is a Compendium of hundreds of operant models, syndromes, patterns, persistent folk ontologies.

42:08 About HAIID

As such, it maps not one but several conceptual models for what Human-AI design interaction may be and might be based upon. It sees HAIID at present as a kind of subset of HCI, but one that arguably will overwhelm and redefine that field, especially as personal AIs are more generally deployed at platform scale.

As I've already insisted, the history of AI and the history of the philosophy of AI are deeply intertwined. One side of that ledger is populated by numerous thought experiments, both canonical and obscure: The imitation game, the Chinese Room, the Paperclip Maximizer, the Three Blue Banana Problem, Samantha's Infidelity, the Driverless Red Trolley, etc.

43:18 The Ledger of Thought Experiments

Among the most notorious new entries may be called simply the "Blake Lemoine" scenario, where the highly evolved tendency to ascribe intentionality to linguistically competent conversants can lead to some unnecessary conclusions. Blaise Aguera Y Arcas and I wrote a piece addressing this episode called, *The Model is the Message*, suggesting that the intelligence there is not quite what Lemoine thought, it was but not quite *not* what he thought it was either.

44:30 The Blake Lemoine Scenario

With many AI interfaces, it would seem that computers have mastered presenting themselves in ways that require almost no additional comprehension for users beyond ingrained social interaction cues.

The history of HCI is in this way a story that shifts from humans having to understand how computers work in order to use them, to computers figuring out how humans work in order to be used by them. Now, language -in its most abstract forms: linguistic reasoning, not only talking and writing- accelerates the latter dramatically and flexibly, even disturbingly, and so draws the practical boundaries of HAIID both deep and wide.

45:16 How Humans Work

As you will see, the capacity for AI to present itself through human social cues is remarkable and, in fact, becomes the interface in and of itself. The shift from HCI to HAIID means a shift from designing click-paths to designing synthetic personalities.

45:47 Synthetic Personalities

Perhaps for the most quantitatively pervasive form of HAIID, is one in which the user doesn't even know AI is there. Things just work. They work and who cares how. However, the forms of HAIID that this project focuses on are those that inspire and extend personal relationships with not only AI but AI persona.

These thought experiments responded to extant AI, and in turn framed and drove further development of the technology. But each was not only a metaphor for what AI "is" but also a scenario for human-AI interaction, and indeed one because of the other. We try to figure out what AI is by figuring out the terms of interacting with it, and to learn how to interact with it by learning what it is: a perfectly understandable approach.

Lately with the rise of LLMs however, there are many new entries to this list, including Sydney's Nervous Breakdown in which, instead of falling in love with his articulate OS as in *Her*, a journalist coaxes a chatbot to perform disturbing feats of abnormal psychology.

What we call personal AIs are central to this and represent a field of tremendous interest, but "Personal" can simply mean AIs that are customized by your personal use of it. They are not necessarily *persona*, but many of the AIs you will use and which use you will be as personalized as your search history, if not your fingerprint.

46:30 Personal AI

The avenue of exploration for this project is then the shift from a form of HCI that is based largely in spatial references -inside versus outside, up and down, over/under- to one that is based in psycho-social metaphors. This is obviously a tremendously powerful shift, but one that comes with all the risks and downsides of human psychology itself.

46:54

Psycho-Social Metaphors

And yet in the most basic form this is not optional. Some projective comprehension in the form of a mental model of what is going on here, in this case, the AI is, what its affordances are, what it is and isn't doing, what and where and why it is and so on. Folk anthologies are not optional.

47:21

Folk Ontologies

Most HAID, we might suppose, will be interaction with, again, in some ways AIs are invisible and boring and unmemorable and yet critical to the reproduction of everyday life, but as said, personal AIs are another matter. They are a kind of AI that is being trained in how you think (like teaching a dog a new trick), but also being trained by your thinking (like carving a rock into an arrowhead). It is a personalization of an external mental model; and is potentially an experience of the self in the third person; if so, how can we not be fascinated?

47:42

Personal Mental Model as a Service

In a moment, I will talk about another project of ours on simulations, but also clearly tie back to the discussion of Personal AIs which are in a sense simulations of us. Or perhaps we are the digital twin of the AI that is working on our behalf. For both, the "sim-to-real" problematics are real and, of course, weird. Perhaps, you are the NPC. Perhaps your shadow is chasing you. Perhaps all personality is a placebo.

48:26

Perhaps, YOU are the NPC

Memory may be the key to any personal alignment worth the name. Perhaps, then, if the uncanny valley is when you are weirded out by something that is but is not quite human, the inverse uncanny valley is when you are much more deeply weirded out by seeing yourself through the eyes of the machinic other. You don't quite recognize what you see but do recognize that what you see is you, but in a way not, but yet it is more real that the version of you that you experience as you. Perhaps you and that newly demystified you will engage in what security teams call "coordinated inauthentic behavior." What is alignment then?

49:00

Inverse Uncanny Valley

The field of HAID is obviously not brand new, in reality it's quite old. But it is new perhaps as a formal disciplinary field of research and design, one that begins as a subset of HCI and may in time come to encompass it. Then if so, does it portend to shift from cognitive psychology of HCI to a renewal of psychoanalysis for HAID? Time will tell.

49:48

The Field of HAID

PART TWO: WHOLE EARTH CODEC

As said, we are clearly quite interested in LLMs, but not just as chabots; we are also deeply interested in a form of what we call *cognitive infrastructures*: the embedding of linguistic competence and hence symbolic reasoning in the inanimate and utterly non-anthropomorphic materials and systems of the world, for which here mind is literally distributed.

50:22

Cognitive Infrastructures

The *Whole Earth Codec* project (by Antikythera studio researchers Chritina Lu, a Deep Mind alum, Dalena Tran and Connor Cook) posits here AI not as a brain in a petri dish but as a synthetic augmentation of the forms of intelligence that emerge in and as complex ecological niches.

50:48

About Whole Earth Codec

The project takes AI as a landscape scale phenomenon, AI in the wild, focusing less on how an AI may align with you or me than how it may align or would align with the wider ecosystem. AI as an inorganic participant in an organic, increasingly self-modeling living world.

51:12

AI as Landscape

Put differently, Whole Earth Codec rethinks the position and application of artificial intelligence as a form of planetary intelligence, and considers potential and necessary conditions for their alignment.

It started by responding to a brief about the quality of data used for foundation models, and by "quality" it was meant both whether the data is any good but also what kind of data it is. Training models that would have global influence on just whatever data happens to be out in the open all but guarantees some degree of suboptimal quality.

51:52

The Qualities of Data

If the most interesting data that could, in theory, contribute to broad-based socially constructive purposes is both private and / or privatized, then other approaches are needed. Parenthetically, techniques like Federated Learning would allow that data to contribute to the reweighing of common models without disclosing underlying values. We could in principle have our cake and keep it private too.

52:46

It's Not About You

But for this project, such a rotation also implies a shift in what kind of data should be *produced*. It ventures that aggregating data about individual human users is only a fraction of what is possible and necessary for planetary intelligence worth the name. It proposes a fundamental de-individuation of computational observation and a focus instead on impersonal, ecological and systemic data.

Instead of centering on individual data with ecological data as a kind of exceptional subsidiary, it inverts this. It posits data about individuals as a specific subset of ecological data which, as should be clear, traces a recurring theme in all our work: culture framed as a function of the planetary, rather than the inverse.

So, the scenario it explores for planetary intelligence is one in which systems of sensing and modeling are global, but importantly it is an observatory looking not outward, but inward. The self-attention of the transformer model is posed as an alternative metaphor to the *panopticon* of Foucault. The positions of observer and observed are less supervisor and supervised than they are mutually recursive.

53:44 From Panopticon to Self-Attention

The model is sensing itself, and thereby the planet is sensing itself: the transformer's self-referentiality is the figure, the allegory for planetary models as well. They call this "folding the gaze".

54:19 Folding the Gaze

The scenario also hinges on the figure and function of multimodality. If we see computation itself as a kind of generic syntax between qualitatively unlike things and actions, then this project locates this generic syntactical function in the sensing and modeling applications of AI as a landscape scale technology.

54:37 Multimodality

The sensing and modeling system is a system for the artificial transduction of planetary phenomena into integrated and recombinant data. Hence Whole Earth Codec.

Multimodality operates both at the level of the kinds of phenomena that are incorporated and artificially mutualized, and in the range of applications and functions to which the system as whole might be directed: mixing and matching inputs and outputs. Multimodal phenomenon, transduced and filtered into a generic syntax, outputted as multimodal application technologies.

In this scenario, planetary intelligence enables planetary ecologies, again, inclusive of human systems, to recompose themselves, because the composition of Whole Earth Codec as a technology for planetary composition enables the emergence of that intelligence. Knowing enables making, but making makes knowing possible.

55:40 Knowing and Making

PART TWO: VIVARIUM

The last project I'll show tonight is called *Vivarium*. Perhaps the philosophy of simulation begins with the beginnings of philosophy itself, in a cave in Greece where Plato and Socrates cultivated a long standing paranoia not just of simulations but of mediated perception and its relation to thought; external and internal simulation in conflict or alignment. There in that cave they've set the foundation not only of what would become a topic for philosophy but perhaps, as I say, the foundational paranoia from which Western philosophy was born.

56:11 Simulationism

The politics of simulation can also be very personal. As you pass through a security gateway, perhaps at an airport, what is under inspection is not only your physical person, but also trace digital personas linked to you but which live in a near-distant shadow city called the Cloud. If the man in the uniform lets you pass, it's because a decision was made according to risk models on those silhouettes of which your physical person is a reflection. Your ears may burn as the infrastructure whispers about your doubles, but it's not just you that's in play.

56:59 The Politics of Personal Simulation

At home and at work, as AI and simulations convene, the designer versus player distinction will collapse from both directions because large AI models and large simulation models will themselves converge, the latter as the interface to the former.

Elsewhere, "scientific" simulations have proposed a different kind of planetary politics based around the frame of climate change that seeks to give political priority and agency to large scale, long duration simulations of macrological processes. It doesn't articulate itself as such, but the core of this approach, the core of climate politics I say, is an attempt to refocus governmental attention from the mediation of voice to the mediation of ecologies, and to make scientifically significant simulations sovereign actors: to make simulations of the future in order govern the present.

57:55 The Politics of Planetary Simulation

Now, *Vivarium* (a project by Antikythera studio researchers Dalena Tran, Christina Lu and Will Freudenheim) deals with the question of sim-to-real rather directly. It poses a platform for collective intelligence that aggregates multiple Toy Worlds into a larger platform of worlds that can be used to train physicalized AI and to aggregate collective data, and thus collective intelligence. It works in various modes: for 1:1 (one human, one AI), between one human and many AI's, many humans and one AI, and perhaps most interestingly for forms of collaborative embodiment, many humans and many AI's.

58:37 Vivarium in Toy Worlds

In practice, I already posited that simulations are an epistemological technology: they are technologies to think with, which in principle makes a philosophy of simulation -a philosophy of things to think with- central to the purpose of any program such as ours.

59:18 Epistemological Technologies

We recognise that simulations are pervasive. Our friends from neuroscience raise the point that simulation is not only a kind of external technology with which intelligence figures out the world, but simulation is how minds have intelligence at all.

The cortical columns of animal brains are constantly predicting what will be next, running through little simulations of the world and the immediate future, resolving them with new inputs and even competing with each other to organize perception and action.

For many computational simulations, their purpose is as a model that reflects reality, such as for climate science or astrophysics. For others, the back and forth is not just mirroring; some simulations not only model their world, but feedback upon what they model both directly and indirectly. We call these “recursive simulations”.

Recursive simulations are those which not only model that reality but which allow us to intervene and interact with it as part of our embodied and intentional experience in a decisive feedback loop. So what are called “digital twins”, perhaps a form of personal AI, are one such dynamic.

1:00:30 Recursive Simulations

As Vivarium shows, many AIs, especially those embodied in the world such as driverless cars, are already trained in Toy World simulations where they can explore more freely, bumping into the walls, until they, like us, learn the best ways to perceive and model and and predict the real world.

For the recursive simulation between the simulation and the real, in some ways the real is the baseline model for simulations and simulations are sometimes the baseline model for the real. Toy Worlds serve as a bounded domain of constrained information exchange and interaction between otherwise unlike and incompatible things and actions.

They are where some AIs learn to navigate the real world by navigating these focused, reductive simulations of their contours. The sim-to-real passage is not just in terms of the implications of specific learned expertises, but also the physical-virtual hybridization as such.

1:01:36 The Sim-to-Real Passage

ML exists in the world, and AI is on its way to become something like a generic solvent, soaked into things and into how they behave. And so the back and forth learning between artificial intelligence and natural intelligence never really stops.

For the project there are, as said, multiple possible combinations of human users, AIs as prostheses, AIs as users, human or humans as prostheses. There are multiple combinations: of embodiment, of agency, of action in and across the simulation, the real and the recursion. Again, not just 1:1, but one to many, and ultimately many to many.

1:02:14 Mutual Prostheticizations

Keep in mind however, for us the AI’s world is a simulation of ours, one we can interact with, for the AI our world is just one part of the omnissimulation that it simply calls reality.

PART TWO: CONCLUSION

While I make some concluding remarks, I will show clips from a fourth project, called Xenoplex, which is on AI and the philosophy of biology, assembly theory and empirical astrobiology (by Antikythera studio researchers Darren Zhu and Connor Cook).

Back to where we began, back to the Stanisław Lem-inspired distinction between existential or epistemological and instrumental faces of technologies.

For AI’s instrumental impact, alignment overfitting is itself a kind of existential risk. As said, capital “A” Alignment is an inadequate practical metaphysics for what AI Orientation implies and demands. Now, I assume most if not all serious alignment researchers would not disagree; if only the journalists, influencers and charismatic mega-critics would follow their lead.

1:03:40 The Existential Risk of Alignment Overfitting

As for AI’s epistemological impact, what will be the ultimate impact of AI on what “we” come to grasp about what we are, how we are, why we are, and the contingencies of that pronoun? That that will be we don’t know and we can’t know. We can’t really anticipate Copernican shifts and traumas in advance; just recall that we didn’t confirm the existence of other galaxies until 1924, or a scientifically confident precise age of the earth until 1953.

1:04:14 The Unanticipatable

We don’t know, but we have to defend the space in which what we will learn will go. We may presume that with regards to AI as an experimental superorganism, one of those likely areas is what is today called “neuroscience” and tomorrow may be called “philosophy”, and vice versa.

So, what are the kinds of questions to be asked that are likely to lead in the direction of epistemic disclosure? There’s likely no wrong answers, but safe bets are that posing fundamental questions about what “life” is and what “intelligence” is and why the “words” that we use may be inadequate signifiers for the range of phenomena that they hope to describe.

1:05:14 It’s the Wrong Words

Even the boundary position between these is unclear, as is the boundary between life and technology and intelligence and technology are already, not just for advanced computers but for anything.

Perhaps life is fundamentally something like “evolutionary autopoiesis through niche technologization”, or perhaps intelligence is, or perhaps both. Biotic systems make use of abiotic systems to replicate themselves as organisms and across generations. They can’t exist without this fundamental technologization of the world. If part of the definition of intelligence is “means agnostic problem solving”, then this niche technologization is at least partially intelligent. The cycle happens not just once or twice, here and there, but everywhere and constantly, over and over, for billions of years. It doesn’t just happen on Earth, it happens by the Earth.

Life / intelligence / technologization is something that the Earth does. And it not only does it, the Earth makes and remakes itself through this process, building scaffolds for the next slightly more complex scaffold, which are incorporated into the next scaffold and so on. It’s why we have an atmosphere and why we have artificial intelligence.

Some planets, at least one, fold themselves over time into forms of matter capable of not only participating in this cycle, but of making abstractions about their own participation in them, and thereby recalibrate them. Human brains are one such form. But they are not necessarily the only form capable of such abstractions and nor are those brains independent of the technical systems of machine sensing and modeling and simulation and prediction that make those abstractions possible. Sapience itself is technological.

More specifically, it’s clear that simulation is, as said, not only part of how animal brains work, how scientific inquiry works, how prediction works; it’s ultimately how the recursion necessary for directed composition works. Simulation as we know it is an advanced coupling of biotic and abiotic systems. It is part of scaffolding.

That is, Biospheres make technospheres that create biospheres that use technospheres to comprehend the whole dynamic.

I think you see where I am heading with this. What is called artificial intelligence is the name for a form of technologization that can occupy more than one position in this cycle. It can be part of the means of technical modeling that humans use to grasp planetary processes, but it can also be the form of intelligence that is doing the grasping. It can be not only a means of planetary sapience but co-constitutive of that sapience as such. That is one way in which the epistemic implications of AI get really interesting and really Weird.

It is then possible to locate AI not just in the reflective shadow of human intelligence, but in the longer arc of intelligence as a planetary phenomenon, and in the emergence of planetary intelligence as such. As suggested, the very definition of these terms is of course put up for grabs not just by philosophy but by what clearly intelligent machinic systems are already doing and by the need to shift the words to the reality. I repeat, *to shift the words to the reality*.

CONCLUSION: AI AND A VIABLE PLANETARITY

I will end with this: there is at present a dangerous disconnect between cosmology and cosmology. By this, I mean that if I go ask my friends in the astrophysics department they will presume I want to know more about Black Holes and Big Bang and curvature of space-time and that kind of stuff. But if I go ask my friends in Anthropology they will presume I want to know about how different cultures imagine eschatology, kinship, and how they think the universe begins accordingly in relation to those.

Now in the Humanities there is, I am sorry to report, significant noise generated around this disconnect. The conclusion drawn adamantly by some is that the abstractions of scientific cosmology must be brought “down to earth”, made to “heal” to the sovereignty of human cultures. I, however, wonder what are the cultures, plural, that can be composed, not just inherited and lived through, that align their ways with the disclosures of the planetary processes with the sapience that make them possible and which are graspable by them? That project is to collectively compose cosmologies adequate to the challenges of long-term planetary viability, not necessarily the reconciliation of that with the diversity of cultural traditions by its subordination.

That is, the ‘disenchantment thesis’ that the modern secularization of the cosmos removed cosmological grounding from culture is wrong; instead it made a *real* cosmology finally possible. Cultural cosmology emerges from the material possibility of thought, and that material possibility of thought emerges from the physical realities that are, in the long run, continuous among humans, even if they exceed the uncertain boundaries of whatever humans are.

So instead of reifying cultural tradition and projecting onto the universe, the better cosmopolitical project for the future is to grasp what is both convergent, because evolutionary, and what is divergent, because human, in the planetarization of civilizations, and to derive abstraction and meaning accordingly.

I hope that the implications for AI alignment, for shifting the words to the reality, are clear. What is at stake for that shift, via AI, is basically everything.

1:05:55

Something That Planets Do

1:07:18

Sapience is Technological

1:07:57

Biospheres Make Technospheres
That Make Biospheres

1:08:32

AI Doing the Grasping

1:09:11

Shift the Words to the Reality

1:09:48

Cosmology vs. Cosmology

1:11:24

The Disenchantment Thesis
Has It Backwards

1:12:00

Convergence and Divergence

1:12:25

Everything