Apocalyptic Technology: AI and the Limits of Science

M. Chirimuuta

28th November 2025
Lakatos Award Ceremony, London School of Economics

0. Introduction

This isn't going to be a lecture about technology bringing about the end of the world. 'Apocalyptic' is meant in the original sense of 'bringing about a revelation', making apparent something that lay concealed. The revelation in question concerns the limits of science. I take inspiration from Emil du Bois Reymond, a German neurophysiologist, one of the first to measure electrical activity in the nerves. He began a public lecture in 1872 with the following words: §

> "JUST as a world-conqueror of ancient times, as he halts for a day in the midst of his victorious career, might long to see the boundaries of the vast territories he has subjugated more clearly defined, so that here he may levy tribute of some nation hitherto exempt, or that there he may discern some natural barrier that cannot be overcome by his horsemen, and which constitutes the true limit of his power, in like manner it will not be out of place, if Natural Science, the world-conqueror of our times, resting as on a festive occasion from her labor, should strive to define the true boundaries of her immense domain."[1]

The lecture in fact became notorious. Controversy ensued from the very suggestion that natural science might have a limit, that there might be some features of the world which scientists will leave forever unexplained. Incidentally, the limit that du Bois Reymond proposed was consciousness, the explanation of how material processes in the brain give rise to our various conscious experiences. To this day, the problem remains unresolved.

I will follow du Bois Reymond in defining the limits of science as the limits of scientists' collective ability to understand the natural world. Like him, I am concerned with the question of whether there are phenomena that we can expect to remain incomprehensible, no matter how far science

[1] Emil du Bois-Reymond (1872) 'Über die Grenzen des Naturerkennens'/ (1874) 'The Limits of our Knowledge of Nature' in *Popular Science Monthly*.

advances. My central claim in this lecture is that AI -- machine learning used as a modelling tool within neuroscience -- has revealed that these limits are tighter than du Bois Reymond had envisaged. Not only consciousness, but seemingly more tractable problems to do with unconscious information processing in the visual system, turn out to have a surprising degree of conceptual and mathematical opacity, due to the brain's remarkable complexity.

Scientists normally have a firm belief in the intelligibility of the universe, a faith that "Everything is figureoutable."[2] § This manifests as a conviction that despite the apparent complexity of natural phenomena, especially biological ones, there is a hidden simplicity which means that some elegant theoretical principles or neat mathematical equations will account for them all, making the once mysterious now comprehensible. § I'll begin my argument with a bit of autobiography, an account of how this belief played a role in my early training in neuroscience, and how technological developments led me to doubt it. But why think that our human capacity for understanding marks the limits of science? The subsequent part of the lecture will present the reasons for this view, followed by an assessment of the current situation in neuroscience: how some researchers are embracing technology without pursuing the traditional scientific aim of offering explanations, while others are seeking new ways to confront the full complexity of the brain while retaining the goal of understanding it. I will finish with some reflections on the role of philosophy in this challenging but exciting new territory.

§

1. Deep Learning and Neuroscience

Though currently my academic home is a department of philosophy, I started out my career as a scientist in a lab. § The research that led to my PhD was a combination of experiments to measure visual thresholds in human subjects (including myself) and programming computer models to simulate the responses of neurons in the early visual system, primary visual cortex, to see if our models could predict the thresholds we actually observed. My supervisor, David Tolhurst, had done seminal work in the 1970s recording directly from visual neurons, seeking to establish the hypothesis that the basic operation of cells in this area is to act as a linear filter of messages originally landing on the retina and passed up via the optic nerve to the brain. § If

[2] "In his modest office steps away from a buzzing refrigerator, Schrag displays an antique microscope—an homage to predecessors who applied painstaking bench science to medicine's endless enigmas. A small sign on his desk reads, 'Everything is figureoutable.' So far, Alzheimer's has been an exception." Piller (2022) *Science* 377:358-363.

the cells are linear it means that you can predict how they will respond to any complicated image or pattern of light sent to the eye – such as the images that you get when you look around this room – just on the basis of how they respond to very simple, artificial stimuli like these bars. If the linearity hypothesis were correct it would also mean that the behaviour of these neurons could be summarised by what is, mathematically, a very simple equation with just a few variables.

§

Already by the 1990s it was clear that there were discrepancies between the data and the linearity hypothesis. In particular, it was found that the activity of any one neuron was influenced by the responses of other neurons around it to an unexpected degree. However, my supervisor and other researchers were aiming to make a few tweaks and additions to the original linear model that would account for these discrepancies and ultimately allow us to predict how the neurons would respond to realistic images, not just the artificial ones. I was in the lab in the early 2000s and we had some degree of success with these adjustments. We published three articles specifically on the problem of extending the model to realistic images[3] but I could tell that David was never satisfied with the results and he carried on tweaking the models in different ways, which he showed me when I came back to visit the lab after I'd left to do research elsewhere and had eventually started my new life as a philosopher of science.

During my time as a student I remember thinking how curious it was that we were attempting to encapsulate the workings of these neurons in mathematical models which were, to be honest, trivially simple. (I hadn't learned any computer programming before starting in the lab, and my education in maths had finished after one year at university, which means that any advanced modelling would have been beyond me.) The thing was that even though I was not doing experiments on the brain myself, I had to read a lot of papers reporting on the responses of these neurons in various kinds of experiments and it struck me that there was a great deal of intricacy here that was hard to pin down. And my supervisor had been studying the visual cortex his whole career and knew far more about these neurons than anyone else in the lab, and that very fact impressed me, how one scientist could devote so much time to one class of cell, and

[3] Tolhurst, D.J., To, M.P.S., Chirimuuta, M., Lovell, P.G., Chua, P.Y. and Troscianko, T. (2010) Magnitude of perceived change in natural images may be linearly proportional to differences in neuronal firing rate. *Seeing and Perceiving,* 23:349-372.
Chirimuuta, M., Clatworthy, P.L. & Tolhurst, D.J. (2003). Coding of the contrasts in natural images by visual cortex (V1) neurons: A Bayesian approach. *Journal of the Optical Society of America A*, 20, 1253-1260.
Clatworthy, P.L., Chirimuuta, M., Lauritzen, J.S. & Tolhurst, D.J. (2003). Coding of the contrasts in natural images by populations of neurons in primary visual cortex (V1). *Vision Research*, 43, 1983-2001.

still not be satisfied that he's figured them out. If one cell is such a challenge, what hope have we for the whole brain? Yet, if the linearity hypothesis were true, there was hope, because in the face of all this apparent complexity there would be one straightforward mathematical principle at play. It shows you how, in our lab, belief in the intelligibility of nature – to be specific, belief in the underlying simplicity of the brain – was so important. Otherwise, we couldn't say why there should be any expectation that some version of a linear model would be the key that would unlock everything we wanted to know about how these cells contribute to our rich and varied experiences.

§

I didn't think too much again about the challenge of building these predictive models until 2018 when I had a sabbatical to begin work on the project that would become *The Brain Abstracted.*[4] I had the idea to revisit primary visual cortex and see how the current results stood. Attending conferences, I'd noticed that many labs had introduced a kind modelling based on deep convolutional networks (similar to AI models used for face recognition) as a means to predict the tough cases where cells are responding to realistic stimuli. The accuracy of predictions was remarkably better than had been achieved with versions of the linear model; at the same time, these models were not based on a linear hypothesis, or any hypothesis about the kind of function that these cells were performing and so the theoretical payoff was unclear. They operated more like a black box, an oracle, that spat out the right answer without explaining to you what was going on. If your motivation, as a scientist, is to figure these cells out, this is unhelpful. I emailed David about this and he shared the sentiment: yes, the new methods work but it means giving up on science, he said.

There was clearly a conundrum:[5] either you build a simple model that makes mathematical sense and gives you understanding of how the neurons respond in the way that they do, but only in a very narrow range of cases; or you can build an immensely more complex, mathematically opaque model that works accurately across the board but doesn't yield understanding of the responses. The goal of science, as I'd grown up with it, was to have both: understanding *and* the ability to make predictions. If you give up on one of these, understanding, then what you're doing isn't really science, it's engineering – that was mine and David's reaction.

§

[4] Chirimuuta, M. (2024) *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience*. Cambridge, MA: MIT Press.

[5] Chirimuuta, M. (2021) "Prediction versus understanding in computationally enhanced neuroscience." *Synthese* 199:767–790

Fortunately, I didn't have to choose between science and engineering because, by then as a philosopher, I had the luxury of sitting back to reflect on the significance of these developments, asking what these new modelling technologies have revealed. For one thing, they seem to suggest that not everything is "figureoutable". There may be some things in nature, like a single cell of a living brain, whose workings are so involved that to predict their responses scientists need to use AI to build models of such mathematical complexity that the models no longer make sense to their users. This would mean that science, with its central goal of understanding the natural world, has a limit; that because of the bounds in how much complexity any one person, or group of people, can make sense of, no matter how intellectually gifted, some questions about the natural world may not have intelligible answers.

For another, it suggested that up until now the working hypothesis of neuroscientists had been that the brain is vastly more simple than it actually is. In retrospect, our linearity hypothesis looked like a piece of wishful thinking. This observation led me to make simplification the central topic of my book: *why had scientists assumed that something as apparently complicated as the brain was fundamentally quite simple? How do experimental methods and controls make neural systems behave as if they were always as simple as we had hoped? How do representational abstractions, such as mathematical models, strip away biological complexity? What are the dangers of over-simplification?* It turned out to be quite a long book and I can't dwell on these other questions now. Instead, I will say more about this uncertain goal of understanding.

§

2. Science and Understanding: The Old Settlement

What the results of the new AI methods in neuroscience suggested to me was that science has a limit. But this conclusion depended on a conception of science in which the goal of understanding the natural world is central and essential. One might object: *why should I be defining the limits of science as the limits of scientists' collective ability to understand the natural world?*

I need to say more about this conception of science. I follow a number of historians and philosophers of science, in a characterisation of the new experimental and theoretical practices that emerged in the 17th century as a hybrid of two previously disconnected activities.[6] § The

[6] John Dewey (1929) *The Quest for Certainty;* Henryk Grossmann (1935) "Die gesellschaftlichen Grundlagen der mechanistischen Philosophie und die Manufaktur" in Freudenthal and McLaughlin (2009); Canguilhem (1937) "Descartes et al technique"; Edgar Zilsel (1942) "The Sociological Roots of

first was natural philosophy, the contemplation of the order of nature. It occurred both in the context of classical antiquity and medieval Christianity, practiced by an educated elite. Understanding of nature was sought for its own sake, as a pure intellectual inquiry, not to be put to any use. The second was mechanical craft, the diverse kinds of practical skills employed by uneducated artisans who needed to produce stuff in order to make a living. As John Dewey, observed, "the distinction was reenforced by social causes. Mechanics were concerned with mechanical arts; they were lower in the social scale" (Dewey 1929, 74).[7]

§

The alliance of natural philosophy and mechanical craft came about with the emergence of a new middle class, as argued by Edmund Zisel, and the development of a machinic view of the universe, as argued by Peter Dear. The natural world came to be understood in the work of 17th century natural philosophers such as Rene Descartes and Robert Boyle as a system closely analogous to a mechanical device in which discrete components interact according to precise physical laws. Such machines are intelligible to their makers, and if the natural world is like that, it too can be understood through a process of reverse engineering, thereby enabling the natural philosopher's task of understanding the natural order. At the same time, mechanical devices are eminently manipulable because they have a transparent and stable causal structure. By treating things as machines, scientists can seek to discern causal structures that can be used to manipulate natural processes and bend them to suit human agendas.[8] Thus the goals of understanding and engineering, previously separate, could be combined. The slogan of Francis Bacon, that knowledge is power, that the attainment of knowledge inherently confers some ability to exert control, seems obvious, unquestionable to us now. However, at the start of the 17th century, the utility of knowledge needed to be argued for, as an alternative to the fruitless intellectual excursions of the Scholastics.

Science"; Paolo Rossi (1962) *I filosofi e le machine;* Peter Dear (2005) "What Is the History of Science the History Of?" *Isis,* (2006) *The Intelligibility of Nature*. See *The Brain Abstracted,* Section 8.2.

[7] See Hadot (1995, chapter 3) on the study of physics as a "spiritual exercise." See Schuhl (1938, 11–12) on this attitude in the writings of Aristotle and Plato.

[8] "…. In this matter I was greatly helped by considering artefacts. For I do not recognize any difference between artefacts and natural bodies except that the operations of artefacts are for the most part performed by mechanisms which are large enough to be easily perceivable by the senses….
Moreover, mechanics is a division or special case of physics, and all the explanations belonging to the former also belong to the latter…..Men who are experienced in dealing with machinery can take a particular machine whose function they know and, by looking at some of its parts, easily form a conjecture about the design of the other parts, which they cannot see.
In the same way I have attempted to consider the observable effects and parts of natural bodies and track down the imperceptible causes and particles which produce them."
(Descartes *Principles of Philosophy* Pt. IV §203 in Cottingham, Stoothoff, Murdoch Trans. 1985 *Philosophical Writings vol 1*. p.288-9)

§

I will now say a little more about the properties of machines so that we can contrast the old picture with the alternative picture of biological complexity that has come to prominence recently.[9] In a machine, the parts do not depend on the whole. The parts pre-exist the whole device, and are put together by the person assembling the machine. The parts are supposed to be stable, not changing with the running of the machine. By envisaging nature as machine-like, it was supposed that natural systems decompose into stable sub-systems. This justifies reductionist investigations focused on low level components, such as cells and molecules, separated from the entire body. In a machine world, things are bounded, modular. It's not the case that everything potentially affects everything else, as was often supposed in the former natural philosophy which, incidentally, was a reason for believing in the efficacy of astrology and magic![10] In a modular world the division of labour in biology is well motivated. It makes sense to study the immune system separately from cardiology, the brain separately from the kidneys. If you tried to study all these systems in detail, at once, things would soon get out of hand. In sum, in the machine world there is an underlying simplicity due to the components being stable and fairly insensitive to context. This lends an intelligibility to the natural world because things turn out to be decomposable into relatively independent modules that can be investigated piecemeal with reductionist strategies.[11]

In case you were wondering what I mean by "understanding", since I haven't yet defined it, some relevant points arise here. Understanding has recently become a flourishing topic within the philosophy of science. According to some, understanding consists in knowing an explanation, where explaining is accounted for in various ways, such as the discovery of biological mechanisms.[12] Another view is that understanding amounts to the possession of certain skills or abilities to perform certain tasks to do with prediction and control of a system.[13] There is a common theme here: all these accounts of understanding relate in some way to the machinic view of nature. The idea that understanding a natural system consists in certain instrumental skills

[9] There is a debate over whether notions of mechanistic explanation in biology really depend on an analogy with classical machines described here. See e.g. Bechtel (2021) 'Living Machines: the extent and limits of the machine metaphor' in Holm and Serban (eds.) *Philosophical Perspectives on the Engineering Approach in Biology,* Routledge.

[10] Mary Hesse (1962) *Force and Fields: A Study of Action at a Distance in the History of Physics.* Ioan Couliano (1987) *Eros and Magic in the Renaissance.*

[11] Even more fundamentally to modern science, the assumptions of stability and decomposability justify the expectation that what is discoveries made in the lab, under artificial conditions, will still be applicable beyond the lab. An unsettling implication of Robert Northcott (2025) *Science for a Fragile World* is that the assumption of stability is much less dependable than previously thought.

[12] Kareem Khalifa (2017) *Understanding, Explanation, and Scientific Knowledge.*

[13] Catherine Elgin. (2017) *True Enough*

stems, ultimately, from the concept of *maker's knowledge*. The builder of a machine knows how it works, and can do things with it, because he or she has put the parts together and is aware of the operating principles. Scientific understanding aspires to that, through the practice of reverse engineering: treating a natural system as if it were an artifact, and trying to discover its operating principles. Likewise, the notion that explanation involves knowing the mechanism of a natural system, or more generally its causal structure, is intuitive only because we think of the system as having a stable network of specific, causal relations, as found in a mechanical device. In addition, there is the shared view that understanding is what leads directly to prediction and control. For instance, pure scientists discover causal relations in a brain area and applied scientists leverage them to create novel interventions, such as cures for a disease.

§

Nowadays, various biologists and philosophers of biology are arguing that the machine conception of the nature is obsolete and needs to be replaced with a *process* view of the living world.[14] A key claim of process biology is that there are no underlying stable parts and causal relations because everything is context dependent and context is itself always changing. What this implies is that ultimately reductionism will fail, that nature is not intelligible through this bottom up method of investigation. It also implies that there are no absolute boundaries around things: the organs within a living body and the organisms within an ecology are not as modular as the machine picture would have us think. There is mutual dependency and relationality across multiple scales. Causal relations are neither fixed nor discrete.

This is a kind of complexity not anticipated on the old machinic worldview. I suggested it was revealed by the inadequacy of traditional models -- models using simplifying assumptions like the linearity hypothesis that ultimately come from reductionism and the machine picture. When put in confrontation with AI models which dispense with those simplifying assumptions, their limitations became clear. § What this also reveals is something historians such as Peter Dear already pointed out: there is a historical contingency in the alliance between natural philosophy and mechanical craft, or more generally, between the activities of seeking to understand the natural world and seeking to manipulate nature through technology. These endeavours came together at a certain point in history, after previously existing as separate spheres of activity; and they could come apart again. With AI we see the predictive power needed for instrumental success in applied science losing its moorings from the understanding sought by pure science. It's reasonable to wonder if the flagship successes of science seen in the last few centuries have

[14] Daniel Nicholson and John Dupré (eds.) *Everything Flows;* Jaeger et al 2023 "An epistemology for democratic citizen science" *Royal Society Open Science*; Nicole Rust *Elusive Cures* (2025 pp. 227-230).

been due to something of a coincidence: the discovery of a delimited range of things for which the predictively powerful theories and models are also intelligible to human scientists. Given these successes, it was assumed that the natural world just is that way: simultaneously intelligible to scientific reason and amenable to technological control. But it could well be that there is nothing in the fabric of the world ensuring that understanding and technological success should accompany one another. Instead, that we should only expect this for the few things in nature that are quite simple -- or rather, simplifiable without a disturbing loss of predictive accuracy. Those quarries of things that are both understandable and controllable have been mined intensively for some time. The divergence of models for prediction and understanding in the 21st century is perhaps the first indication that a resource, once taken to be unlimited, is depletable after all. And if this raw material for science—things subjectable simultaneously to scientific understanding and manipulation—is limited and nonrenewable, it means that science itself has a limit. At least, we see that the old settlement is no longer tenable. In the next part of the lecture we'll consider some developments that are occurring in neuroscience, given the breakdown of the old settlement.

§

3. The New Prospects

What happens when scientists give up on the assumption of underlying simplicity within nature, the assumption which justifies use of widespread simplifications such as linear models and reductionist methods? Specifically, what happens when neuroscientists give up on the idea that the brain is not more simple than it seems, that it is not amenable to elegant, unifying and explanatory theories of the sort prized in physics?

§

I will now describe three responses to the new situation: first, that neuroscience becomes neuro-*engineering*; second, that neuroscience continues to pursue its twin goals of understanding and intervention, but without the assumption of underlying simplicity; third, that natural philosophy re-emerges as an endeavour divorced from the goal of intervention.

§

3.1) Neuro-engineering Replaces Neuroscience

Under the old settlement, science was formed of an alliance between natural philosophy and technology, meaning that it inherited the twin goals of understanding the natural world and seeking to control it. Given the breakdown of the old settlement, one option is to drop the goal

of understanding and let science develop into a pure technological endeavour, aided by AI and other tools for prediction and control.

Back in 2008, half a decade before the start of the current wave of AI through deep learning, Chris Anderson, editor of *Wired* magazine prophesised that the data mining methods perfected by Google would show their effectiveness in the sciences, where accumulation of data had outpaced theoretical innovation.

"[M]odels, ….." Anderson wrote, "are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works." He continued, "the more we learn about biology, the further we find ourselves from a model that can explain it. [However,] There is now a better way. …. We can stop looking for [explanatory] models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

Anderson has certainly been criticised for over-stating the case for hypothesis free models and theories – science as data mining, without scientific expertise.[15] For one thing, successful use of these methods depends on high quality data, and without scientific expertise, data quality cannot be evaluated. The central law of computing is as relevant as ever: *garbage in, garbage out*.

That said, Anderson hit upon a trend that would occur in some form:[16] not that computer engineers would replace scientists, but that scientists would take up the methods of engineers. § An example of this tendency comes from Jim di Carlo's lab at MIT. This group was amongst those leading the introduction of deep learning in visual neuroscience, as I discussed earlier. In one study from 2019, they built an artificial neural network (ANN) model of neuronal responses in area V4 of the visual cortex and used this model to devise stimuli that would maximally excite those neurons, driving them to produce more action potentials than when exposed to any other stimulus. One of the purposes of the study was actually to address criticisms that their previous

[15] Wolfgang Pietsch (2021) *Big Data* p.62 ff.

[16] Anderson's case has been restated in a more compelling way by Evans and Duede (2025) 'After Science' *Science:* "With the emergence of AI in science, we are witnessing the prelude to a curious inversion – our human ability to instrumentally control nature is beginning to outpace human understanding of nature, and in some instances, appears possible without understanding at all. With rapid adoption of AI across all scientific disciplines, what does this mean for the future of scientific inquiry? And what comes after science?"

ANN models of the visual cortex were impressive gadgets that did not yield understanding of the brain (Bashivan et al. 2019, 1). In this study, by showing that their model can be used "to …. control the brain", they believed, they had, to quote, "a practical test of useful, causal 'understanding'" (11). This is a very interesting remark because in effect it lays down a redefinition of scientific understanding. From a sense of comprehension or ability to grasp the principles of a thing they switch to a purely pragmatic criterion: a *system is understood if and when it can be controlled*.

Previously we saw that scientific understanding has been tied to instrumentality as maker's knowledge – the understanding available to the designer of a system, in virtue of its operating according to her design. The thing about artificial neural networks is that they are self-organising systems that learn the solution by following a training algorithm. Therefore, the maker of the ANN does not know the principles of the solution embedded in the trained network. As with actual neural systems, the workings of trained artificial neural networks are quite opaque.

Two neuroscientists commenting on the di Carlo project make some interesting observations, comparing this new notion of understanding with the former one in which understanding required the *reduction* of a system into parts that are simpler and hence more intelligible. Batista and Kording write that,

> "The ANN approach as applied by Bashivan, Kar, and DiCarlo advances a different conception of 'understanding': we understand a system when we can control it. Most of us understand our cars in this latter sense (predictability and controllability) without really understanding them in the former (reductive) sense. The authors, by showing that they can steer neural populations to preselected states, have demonstrated that they indeed understand the visual system, in the sense of controllability. We wonder, is the reductive notion of understanding even possible for a system as complex as the brain? Or might the 'I can control it' notion of understanding actually be a more effective and relevant way forward for neuroscience?" (Batista and Kording 2019:564 on Bashivan et al)

Still, one point of continuity between the old settlement and this tech driven trend is that they both employ the machine conception of a natural system. Neuroscientists such as di Carlo embrace the idea that the brain literally is a computing machine, executing algorithms like the

ones running in his ANN servers. The difference is that with the new kinds of machines, even their makers don't know how they really work.[17]

§

More recently, neuroscientists have taken up the idea of *foundation models* from the tech sector. This is the kind of modelling strategy behind large language models such as chat GPT, one that requires exhaustive amounts of data. A group from Andreas Tolias's lab showed that by training a foundation model on data of about 100,000 neurons recorded from the visual cortex of 14 different mice, this approach achieved an unprecedented level of accuracy, when challenged to predict the responses of neurons in another group of mice, to novel video stimuli (Wang et al 2025). Within psychology, actual large language models, tweaked to predict the results of behavioural tasks given to human participants in the lab, have been shown to outperform most traditional models in the field, in terms of predictive accuracy (Binz et al 2025).[18]

Amongst scientists, this development has of course raised the question of what can be learned and understood about human cognition by having an AI that can reproduce human like results without revealing the processing steps which led to the outputs. Stanford psychologist, Russell Poldrack, has offered an answer here.[19] § His view is that the predictive power of a model is not in tension with its ability to provide understanding, that there is not a trade-off between prediction and understanding, as I, for one, have argued, because the concepts of prediction and understanding are linked: both involve data *compression*. …….

The issue is just that the data compressions that make intuitive sense to us humans are low dimensional and too lossy (if the dataset has multiple, complex interactions), whereas AI achieves compressions that are high dimensional and not too lossy, which is why they are able to achieve predictive power where traditional models, the low dimensional ones that are intuitive to us,

---

[17] Bashivan, Pouya, Kohitij Kar, and James J. DiCarlo (2019). "Neural Population Control via Deep Image Synthesis." *Science*; Aaron Batista and Konrad Kording (2019) "A Deep Dive to Illuminate V4 Neurons" *Trends in Neurosciences*.

[18] Eric Y. Wang et al. (2025) "Foundation model of neural activity predicts response to new stimulus types" *Nature;* Marcel Binz et al. (2025) "A foundation model to predict and capture human cognition" *Nature:*

"Here we introduce Centaur, a computational model that can predict and simulate human behaviour in any experiment expressible in natural language. We derived Centaur by fine-tuning a state-of-the-art language model on a large-scale dataset called Psych-101. Psych-101 has an unprecedented scale, covering trial-by-trial data from more than 60,000 participants performing in excess of 10,000,000 choices in 160 experiments. Centaur not only captures the behaviour of held-out participants better than existing cognitive models, but it also generalizes to previously unseen cover stories, structural task modifications and entirely new domains." (LLM was Llama, pretrained by Meta AI)

[19] Presentation at SPAN – Society for Philosophy and Neuroscience, May 2025: "Prediction, compression, and understanding".

have failed. Moreover, Poldrack argues that AIs such LLMs are not as opaque as we feared: with some probing, some of their workings can be interpreted.

§ …….

On this last point, I think we have a situation of glass half empty/half full. Indisputably, AIs are more opaque than the models traditionally built by scientists. The fact that they are not completely opaque black boxes, that they can be reverse engineered to some extent, so that model users gain some clues as to the representations and processing steps that occur within a trained model, gives the optimist like Poldrack cause to say, well at least they offer something, by way of helping us to understand the systems we're predicting. That they are far less intelligible than the previous standard of explanatory models gives the pessimist like me reason to say that a trade-off does obtain: the greater predictive power of these models is bought at the cost of having a model that makes much less sense to its users. There will always be an opacity to these technologies, an ability to generate answers without anyone fully knowing how, since otherwise they would not be useful as an alternative to unaided human cognition. If really transparent they would only ever produce the answers that a user could, in principle, know in advance, from knowing how the device works. As applied physicist Kevin Kelly pointed out, "the power of a technology is proportional to its inherent out-of-controlness, its inherent ability to surprise and be generative. In fact, unless we can worry about a technology, it is not revolutionary enough" (quoted in Dupuy 2009:xii)[20].

My conclusion, therefore, is that the faster neuroscience advances with AI, the further the traditional goal of scientific understanding will be left behind. This accelerationist trend gives cause for concern, for it allows intervention to be pursued without theoretical understanding. We've learned through bitter experience that unintended consequences always follow interventions, even when due theoretical diligence has been done; how much more so when neuro and bioengineering operates unchecked by the aim of figuring out what is actually going on. That could be apocalyptic in the usual sense of the word.

§

3.2) The New settlement

Within neuroscience and other branches of biology we also see an alternative direction developing in response to the breakdown of the old settlement. I call this one the new settlement

[20] Jean-Pierre Dupuy. (2009) *On the Origins of Cognitive Science.*

because it retains the twin goals of understanding and instrumentality, and the notion that pure science will lead to applied science, but it drops reductionism and the machine picture, or at least that is the intent.[21] With this, it also gives up the presumption of an underlying simplicity to the living world, in favour of a much more dynamic and complex view of things. The challenge is to pursue the goal of understanding nonetheless.[22]

Philip Ball has written about this new path in the popular science book *How Life Works,* beginning with the admission that previous molecular and genetic approaches in biology have been guilty of over-simplification*:*

> "just about all the neat stories that researchers routinely tell about how living cells work are incomplete, flawed, or just totally mistaken.
>
> All the same, I believe we *can* do better. I will show how research in molecular and cell biology over the past several years has painted a richer and much more astonishing picture than that bleak and obsolete mechanical metaphor. The picture does at times appear fantastically baroque and perplexing, but in the end it takes the burden of control off the shoulders of the genome, relying instead on principles and processes of self-organization ~~that, precisely because they have no need of tight genetic guidance, avoid the fragility that would engender~~." (Ball 2023:4)[23]

§

Similarly, the neuroscientist Nicole Rust has published a book this year placing blame for the lack of advances in translational neuroscience – its failure to discover cures for neurodegenerative conditions such as Parkinson's and Alzheimer's – at the door of reductionism and its simplistic, linear model of causation. […….] The old picture is to be replaced with a view of the brain as a *complex adaptive dynamical system*. As with Ball's reference to "self-organization", a new suite of concepts is borrowed from complexity science. Instead of linear causation like a domino chain, we have a dense causal web where, due to the ubiquity of feedback, it becomes unfeasible to disentangle causes from effects. § Processes contributing to

[21] As put into slogans: Johannes Jaeger: "Beyond the age of the machines" (https://www.expandingpossibilities.org/0-introduction.html)
Philip Ball: "The End of the Machine" (2023, chapter 1).

[22] The reaction against the machine-based view of life is not as novel as sometimes made out by its recent proponents. On the "oscillation" in the history of biology, between mechanistic and anti-mechanistic views of life, see Georges Canguilhem (1955) *La Formation du Concept de Réflexe aux XVIIe et XVIIIe Siècles;* also Anne Harrington (1996). *Reenchanted Science: Holism in German Culture from Wilhelm II to Hitler.*

[23] Philip Ball (2023) *How Life Works*

pathology occur at multiple scales beyond the molecular level, and interact with one another across scales. (Rust p.110).[24]

With this picture in place, we can appreciate why expectations for neural therapies have not been met, and better see the inherent difficulty of the problem. Traditional "bench to bedside" methods approached a disease like Alzheimer's at one level, the lowest level here depicted: genetic susceptibility leading to abnormalities of proteins in the brain, this being the supposed cause of neuronal death. Enormous sums of money were spent on developing a treatment which very effectively removes the suspect protein, Amyloid beta, but this turned out to have negligible clinical impact on cognitive decline.[25]

On the new approach outlined by Rust, we must conceive of Alzheimer's as a multi-level pathology, with causal loops going through the environment, potentially impacting processes at all the other levels. But we cannot bring the environment into the lab. Previously, Alzheimer's researchers studied model systems in labs, abstracting away from complexity due to environmental interactions, generating tractable problems. But their advances were lost in translation because the actual disease occurs in a human brain housed in a body, belonging to a person who belongs to a family and lives in a society.

§

The continuity between the old and the new settlements consists in them both counting on there being a path from understanding to instrumentality, from pure science to applied science. As Rust writes, "A fruitful route to the end goal of controlling a complex system begins by understanding it (ideally with great detail)" (p.237). An important difference lies in the scale of the challenge that comes with the attempt to understand a complex adaptive system. Because reductionist methods are insufficient, and it is unwise to ignore contextual variables, it becomes necessary to collect data, exhaustively, at multiple levels, casting the net as widely as possible.

---

[24] Nicole Rust (2025) *Elusive Cures.* "To move forward in a manner that is impactful for understanding brain dysfunction, we will need to shift forward appreciating that the 'things' to be explained need to be defined in ways that require feedback. If we perserverate on problem that are too static and too simple, Occam's razor will continually draw us back to domino chains. Understandably, we spent a few decades applying that approach to focus on tractable problems and solutions. Moving forward, if we want brain research to be impactful, we must move beyond those easier problems and tackle harder ones." (p.226-227)

[25] Shi, J., Sabbagh, M.N., and Vellas, B. (2020). Alzheimer's disease beyond amyloid: strategies for future therapeutic interventions. BMJ *371*, m3684. https://doi.org/10.1136/bmj.m3684.

van Dyck, C.H., Swanson, C.J., Aisen, P., Bateman, R.J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., et al. (2023). Lecane- mab in early Alzheimer's disease. N. Engl. J. Med. *388*, 9–21. https://doi. org/10.1056/NEJMoa2212948.

Rust is optimistic about the prospects for achieving this because technologies are now available for simultaneous recording in the brain, at multiple levels. However, there is the threat of the data deluge: masses and masses of incomprehensible results. Again, Rust is optimistic because of technological advances in machine learning. Algorithms can mine the data for patterns. Recurrent artificial neural networks, which are models of adaptive dynamical systems, can be trained to perform cognitive tasks and reverse engineered to yield insights into the brain data. (p.237-8)

§

In a paper cited by Rust, Rollo and co-authors[26] offer a blueprint for a multi-scale dynamical systems approach to Alzheimer's. It is worth mentioning their four principles….

1. ["We must embrace ignorance and uncertainty"] They go on to say: "A complex non-linear system can, and usually does, behave in a way that cannot be predicted from the behavior of its parts." In other words – reductionism fails

2. ["We will never know that our description of the system is complete and must continue to challenge consensus."] This is because the characterisation of 'the system' is always a work in progress, due to failure of modularity.

3. ["*we cannot design interventions by focusing on part of the network*."] Again, this relates to the failure of the assumption of modularity.

4. ["An ecosystem approach ~~that maintains a diversity of hypotheses, builds trust, and shares results is essential for progress~~"] Here, the endorsement is for interdisciplinarity and pluralism, which is not a new idea when it comes to dealing with complex multiscale phenomena.

§

It's time to offer some assessment of the new settlement. Firstly, it remains to be seen if it can live up to its ambitions of providing a new path to therapeutic control via understanding of these unfathomably complex processes. Rust is upfront about the fact that this approach is under construction and has not yet born therapeutic fruit (Rust P.230ff).[27] As we saw just now in the

[26] Jennifer Rollo, John Crawford, and John Hardy (2023) "A dynamical systems approach for multiscale synthesis of Alzheimer's pathogenesis" *Neuron*.

[27] Cf. "Whether this strategy [of treating cancer] will pan out remains to be seen, but it's a good example of the philosophy of treating disease at the right level of intervention: matching the solution to the problem. If cancer is at root a matter of cells falling into the 'wrong' state – the wrong basin of attraction – perhaps the real goal is to get them back out again. That might have more in common with the kind of cell-state

fourth principle, alongside the scientific challenge of collecting and modelling the data, there is the sociological challenge of reorganising research to enable the different kinds of collaboration required by the approach.

Secondly, we can ask whether a new notion of scientific understanding is viable without reductionism and the machine picture, or if what is meant by these researchers as 'understanding' is ultimately bound up with the old approach and its ideal of maker's knowledge. For it is not obvious what understanding now means. As Rust admits (p.111), the meaning of scientific understanding is "elusive". Models of multi-level dynamical systems models are not intelligible in the sense first articulated by Werner Heisenberg and more recently by philosopher Henk de Regt:[28] we cannot say, qualitatively, how altering parameters and variables will affect the behaviour of the model system without actually running the simulation. As Rust puts it, these models "defy intuition".

It may be unavoidable that understanding requires simplification, that what scientists are aiming for when they seek understanding is a neat, intuitive picture of how a system behaves. To borrow Poldrack's term, they seek a compression of the original, messy state of affairs into a low dimensional representation in which a very small number of factors can be shown to generate the key phenomena.[29] The following passage from Rollo and co-authors indicates that scientists pursuing the new approach are not in fact giving up on the dream of discovering simplicity underneath all the complexity:

> "This brings us to one of the outstanding challenges of AD: among the almost infinite level of detail that is possible to measure, is it possible to identify the essential detail that is needed to design interventions that control the system-level behavior? There are some clues to the answer from dynamical systems theory, which give us some hope that this question can be answered. For example, widely different systems from earthquakes and economics to ecosystems and cell networks display the same kind of ''critical behavior'' where abrupt changes in the state of the systems can be brought about by small changes in a few or a single control variable. There is some recent evidence that this kind of behavior also occurs in the brain. Because these behaviors are observed in widely different systems, they are only weakly dependent on the details of the systems, suggesting a path to parsimony. This is an attractive insight and the approach we

engineering involved in stem-cell research than with developing drugs against molecular targets. It's about redirecting life itself to new destinations." (Ball 2023 p.411)

[28] Henk De Regt (2017). *Understanding scientific understanding*.

[29] Rust (2025 p.226): understanding requires abstraction.

advocate for in this paper is focused precisely on determining this pathway to parsimony in AD." Rollo et al 2023:2130-31

This passage is very revealing. For one thing, the scientists are seeking one or a small number of "control variables" whose effects on the system are stable and clearly traceable, and can serve as handles for intervening on it. Just as under the old settlement, understanding is linked to instrumentality because both involve the tracing of a causal structure, one that is useful and intelligible at the same time. Rollo et al argue that even for a complex dynamical system, a set of simple causal relationships might be ascertainable because, in effect, the system simplifies itself. That is, most of the details of the system turn out to be irrelevant to the global behaviour of interest (e.g. the transition from health to disease) because there are high-level patterns that occur in a range of physically very different dynamical systems. Researchers can focus on these patterns and ignore much of the nitty gritty detail.

§

However, Nicole Rust (p.192) points out that more often than not, hopes for extremely general models of complex dynamical systems have not been met. The devil is often in the detail and modellers have to dig around to find out which details and variables matter most to determining the behaviour of different kinds of systems. This relates to a general remark, which is that the dynamical systems framework originated in physics, along with most of the 'coarse graining', simplifying procedures used to discover high level patterns of the sort Rollo et al discuss. In physics, the evidence for separation of scales, the independence of processes at different levels, is much stronger: that is why people can build bridges using Newton's laws, ignoring quantum level effects. In biology, in neuroscience, molecular level processes constantly interact with macro-scale ones (hence pharmacology). Particular methods for coarse graining, for abstracting away from lower level details of a system, such as the normalisation group, make assumptions about the homogeneity of the lower level entities which are reasonable in physics (dealing with atoms), but not in biology, since living cells are highly heterogeneous.[30]

Another general worry is that the central concepts and techniques of complexity science are quite old (nearly as old as me) and have not produced the breakthroughs elsewhere that people might have hoped for.[31] The optimism about these methods in neuroscience could be due to the

[30] Timothy Lillicrap and Konrad Kording. (2019) "What Does It Mean to Understand a Neural Network?" https://arxiv.org/abs/1907.06374.

[31] John Horgan (1995) "From Complexity to Perplexity" *Scientific American*.

novelty of their use in this field. Even though the brain is often called the most complex object known, neuroscience is actually quite late to the complexity science party.

§

3.3) Natural philosophy

If the new settlement doesn't work out, we can expect that understanding and intstrumentality will go their separate ways, no longer harnessed together in the marriage that has defined modern science. We can ask what would be left of the project of understanding by itself, no longer tethered to instrumental aims. Would there be a return to natural philosophy as it once was?

In this lecture we have seen how scientists have sought simplicity in nature by conceiving it as machine-like. In many domains, this has brought astounding instrumental power: electronics, explosives, antibiotics, molecular diagnostics. However, as scientists have pushed these same simplifying methods into the domains of more and more complex systems, such as the brain, their limitations have shown. In order to achieve precision and theoretical clarity, scientists need to put boundaries around things because they cannot model everything, everywhere, all at once. But if, to quote the biologist Richard Levins, quoting Hegel, "the true is the whole"[32] – if it is the case that in the domain of complexity everything, everywhere is feverously interactive – then we might be restricted to acknowledging the truth of this situation, without being in a position to master it.

And so we would be brought full circle to a kind of knowledge that can appear paradoxical, even counterfeit, after such a long habituation to the Baconian maxim that knowledge is power: we would have attained some useless knowledge. Yet, there is a consolation here that I mention in the final chapter of my book. I argue that the mind-body problem as we know it since the 17th century is a product of the same demand for modularity, for bounded domains of investigation, that characterises the modern scientific project. This was a point often made by pragmatist philosophers such as James and Dewey: in order for there to be a rigorous science of physics, qualitative, mental properties had to be excluded from the domain of the physical.[33]

---

[32] Richard Levins (2006) "Strategies of abstraction" *Biology and Philosophy.*

[33] William James (1890/1950) *Principles of Psychology:*

'The desire on the part of men educated in laboratories not to have their physical reasonings mixed up with such incommensurable factors as feelings is certainly very strong. I have heard a most intelligent biologist say: "It is high time for scientific men to protest against the recognition of any such thing as conscious- ness in a scientific investigation." In a word, feeling constitutes the "unscientific" half of existence, and anyone who enjoys calling himself a "scientist" will be too happy to purchase an

Finding a place for them in the natural world then became rather difficult. But if we accept that the notion of the purely physical is a simplification, an idealisation, we can say that as a matter of fact, but not a fact susceptible to rigorous scientific elaboration, the mental and physical are inherently linked, or at least not radically different. The dualist intuition that prompts the mind-body problem can be dissolved.

Back to where we started. Du Bois Reymond argued that explanation of the mind body connection was beyond the limits of science. From the reanimated perspective of natural philosophy, we could instead take the apparent disconnection to be an artifact of the scientific demand for theoretical and conceptual clarity. We can therefore know that mind and body are not disconnected and relieve ourselves of the unsettling feeling that science ought to provide an explanation of their connection, though frustratingly it cannot. A piece of philosophical knowledge that is, admittedly, unactionable – useless.

§

4. Summing Up

In this last slide I summarise our journey towards and perhaps beyond the limits of science. Here, in the territory of the old settlement, instrumentality and understanding worked hand in hand. The systems investigated were simple enough that low dimensional representations of their causal structures were predictively accurate and hence instrumentally effective.
To borrow Poldrack's term, the compressions here executed by abstract models make sense to human scientists and tie prediction and understanding together.

Beyond that, we have the territory of highly complex processes. Over here machine learning algorithms are generating high dimensional compressions of the data that do not make too much sense to their users. Prediction and understanding are pulling apart. Engineering is taking off on its own.

Over there, scientists are sticking to the old template of seeking instrumental control by way of understanding the system first. It remains to be seen if models that make sense to the scientists

---

untrammelled homogeneity of terms in the studies of his predilection, at the slight cost of admitting a dualism which, in the same breath that it allows to mind an independent status of being, banishes it to a limbo of causal inertness, from whence no intrusion or interruption on its part need ever be feared.' (pp.134–135). See *The Brain Abstracted,* Section 10.4.

can be anything other than drastic simplifications of highly complex systems, compressions that are too low dimensional and too lossy to be predictively accurate and instrumentally successful.

Over here a solitary natural philosopher contemplates this buzzing, seething, disorderly order of nature without seeking to tame it. The result could be therapeutic. As Wittgenstein said, the task of philosophy is to leave everything as it is.

Thank you.