Human Arenas
https://doi.org/10.1007/s42087-024-00415-4

RESEARCH

From the Reflex Machine to the Changeable Brain

Mazviita Chirimuuta[1]

Received: 29 January 2024 / Revised: 31 January 2024 / Accepted: 27 April 2024

Abstract
The brain is exceedingly complex. Neuroscientists have used the tried and tested simplification methods of physics, such as mathematical idealization, in order to deal with neural complexity. This paper queries the appropriateness of certain physics-derived strategies within this biological science by presenting a historical case study (the reflex theory) in which scientists fell into the trap of over-simplifying the brain and nervous system. It then considers whether more recent computational and dynamical systems approaches in neuroscience are also at risk of abstracting away from critical features of neural functionality related to the inherent changeability of those systems.

Keywords Philosophy of neuroscience · History of neuroscience · Behaviorism · B.F. Skinner · Ivan Pavlov · Charles Sherrington · Dynamical Systems Theory · Computationalism in neuroscience

Introduction: Querying Simplicity

This article is a re-elaboration of the talk I gave at the workshop on physics modelling of thought and partly based on a new book of mine entitled *The Brain Abstracted* (Chirimuuta, 2024). As I was asked to feature in the session on historical and critical perspectives, I decided to do a bit of both: some history and a few provocations at the end. The book is about the role of simplification in the past and the present of neuroscience: it deals with how neuroscientists have confronted brain complexity and tried to find their way through the tangled forest of many different neural systems. Many of the ways that neuroscientists have tried to simplify have, of course, been inspired by physics, which means there has been plenty of transfer from the methods, community, and conceptual schemes of physics into neuroscience, precisely because these offered a set of ways to simplify the brain.

✉ Mazviita Chirimuuta
m.chirimuuta@ed.ac.uk

[1] University of Edinburgh, Edinburgh, UK

Published online: 11 May 2024

I will begin by noting how in the background philosophy of physics there is a belief in the fundamental simplicity of the universe. And you can see that with the importation from physics into neuroscience and other parts of biology, life scientists have followed physicists in trying to seek out simplicity in what are the manifestly complex systems that they are investigating. In the following I am going to talk through an example of this process which occurred in the early 20th century, resulting in a general theory of how the brain and central nervous system works. This theory posited that the whole nervous system is decomposable into simple reflexes and it is not, in hindsight, a very good theory of how the brain and nervous system works. I will argue that it is that very seeking of simplicity that perhaps misled the scientists. Seeking simplicity may not always be a good thing and my somewhat provocative claim here is that maybe this importation of physics into biological sciences has its problems.

Following my discussion of the reflex theory, I will be talking about the transition into the computational theory of the brain that happened in the mid 20th century and then make a few remarks also about current use of dynamical systems theory in neuroscience.

As an example of how physics prizes simplicity, I report a quotation which is attributed to Einstein: "I have deep faith that the principle of the universe will be beautiful and simple." The way that this desire for simplicity is manifest in physics is in the drive for fundamental theories which are elegant, that is, expressed by a handful of mathematical symbols combined into equations which are thought to allow one to predict and explain whole realms of processes occurring in nature (Boge et al., 2023). The core idea is that underneath all of the complexities of the universe there is a simple set of laws of nature at the heart of it. But is this more than an article of faith? When I was looking into this, I was working on the introduction for my book and I found it had some actually quite deep theological roots. Einstein in his later career thought that the principle of the simplicity of nature had instantiated itself in his own empirical research, so he thought there was at least some experiential basis in his own career for asserting it. What I want to point out here is that the value of simplicity in science is not incontestable, especially in biology where mathematization and reduction of complex systems to a few simple equations has not been as productive an approach as in physics. We should therefore query whether it is always valuable and productive for biologists to look for the simplest ways of representing their systems.

Lessons From the Past

As a guidance to our reflection, let us ask ourselves the following questions: historically, have the simpler theories turned out to be better? Have simpler models turned out to be more fruitful? Are there significant historical cases of simplification-gone-wrong? From these we can extract potential lessons for contemporary science.

Reflex Theory

The reflex theory of the brain and nervous system is the one that I think we can take as a case of simplification gone wrong. When we look back at this theory, it does seem, given what we know about it now, really to be an oversimplification. Yet, it was very compelling to people at the time precisely because it did offer a grand unified theory of the brain, where all

of the processes in the brain were decomposable into simple reflexes. This approach to the nervous system began in the late 19th century, after the discovery of the simple spinal reflex arc. There were the first murmurings of it within the last quarter of the 19th century and then in the first couple of decades of the 20th century it became a dominant view. Postponing for a moment the explanation of the details of the theory,[1] let us read a passage which conveys the prestige of this theory at the time, one hundred years ago:

> The concept of reflex action has played, in the 19th and first twenty-five years of the 20th century, a dominating role, comparable, perhaps, to the influence of the Newtonian hypotheses in physics. (Fearing, 1930, p. 4).

These words were written by Franklin Fearing just after the heyday of the theory, in 1930, in his historical and somewhat critical overview of the theory, adding that the reflex theory had the status of Newtonian physics "[f]or those sciences primarily devoted to the study of integrated responses of living organisms,"--i.e., the science of how organisms have coherent responses to their environment, mediated by nervous connections. Charles Sherrington, whose name is one of the grand names in the history of neuroscience, was also committed to the reflex theory. His magnum opus, *The Integrative Action of the Nervous System* describes in detail simple and complex reflexes based on his own physiological research on the motor system.

Newtonian mechanics is now, since the Einsteinian and quantum revolutions, considered a false theory in its details, but still highly useful, finding application in the domains of macroscopic medium size objects. Unlike Newtonian mechanics, the reflex theory of the nervous system is no longer considered a useful way of thinking about the brain and nervous system, nor taught in school and universities. To be sure, in today's textbooks you still have depictions of the sensory motor reflex arc, as that bit of anatomy and physiology is, of course, still current, but the idea that all of the actions of the nervous system are fundamentally decomposable into simple reflexes is not something that neuroscientists today would be committed to. As far as I gather, it is not even taught much in terms of the history of the field, within the kind of historical training that neuroscientists themselves obtain. I only came across the theory many years after my own study of neuroscience, when I was looking back in history for more philosophical purposes.

I now want to give you a bit more of a sense of what the reflex theory, this 'reflexology', as it was known back then, actually was. The core concept is that there exists something analogous to atoms in the nervous system, a sort of elementary unit in behaviour. The role of the atoms or elements in the history of the physical sciences has been to serve as the simple, homogeneous, stable units which underlie the complex phenomena that are observable. Physiologists like Jacques Loeb and psychologists like Ivan Pavlov took the idea that in order for biology and neurophysiology to advance, there needs to be a discovery of what the elementary components of their systems are:

> The understanding of complicated phenomena depends upon an analysis by which they are resolved into their simple elementary components. If we ask what the elementary components are in the physiology of the central nervous system, our attention is directed to a class of processes which are called reflexes. (Loeb, 1900, p. 1)

[1] see Chirimuuta (2021).

When Pavlov (1927/1960, 8) wrote that the reflexes are, "the elemental units in the mechanism of perpetual equilibration," he meant by this the way that the organism is adapting itself, keeping itself in balance with its environment, could only be understood by decomposition into reflexes. One thing that you'll notice here is that the reflex theory is working both at the level of behaviour and neurophysiology. Pavlov became famous for his experiments on learning in dogs, classified as research in psychology, but he thought there was a direct correspondence with the physiology of the nervous system, where you can look for the actual neural basis of conditioned and unconditioned reflexes. For him these two levels of the reflex theory were quite directly connected.

In terms of how physics is playing a role here, one thing to think about is the notion of analysis in classical mechanics, which is being imported into the way that people think about the brain and behaviour. The idea of elements and chemistry give only a very loose analogy to what is being sought here with the elementary reflex–as Loeb says "simple elementary components". A more direct inspiration is from the idea of decomposition and analysis in classical mechanics. And in terms of a philosophy of science in the background animating this, Loeb (1912, 58) was quite aware of it, stating that, "it's better for the progress of science to derive the more complex phenomena from simpler components than to the contrary." Looking at successful science in other fields, his view was that there is a trajectory of going from simplicity to complexity–decomposing the system into some simpler units, getting a full understanding of those, and then seeing how the simpler units build up to more complex phenomena. That is what he was taking from other branches of science and seeing how the same strategy can work in neurophysiology. Noteworthy in this context is the title of his book, *The Mechanistic Conception of Life*. Loeb was part of a movement of biologists known as 'physicalists' and 'mechanists', who were standing against the schools of biology that saw living organisms as something quite *sui generis* and not susceptible to the methods of the physical sciences. In contrast to them, he believed that progress in biology would be achieved by thinking of organisms simply as mechanisms and bringing the methods of the physical sciences into biology.[2]

The method of simplification by reduction or analysis has been commented upon recently in the history and philosophy of science by Bechtel and Richardson (2010) in their *Discovering Complexity*. An interesting early source is a work by the physicist Percy Bridgman, well known for his operationalism in the philosophy of science. His observation on physics, not biology, is that:

> A conviction of the significance of microscopic analysis has many features in common with the usual conviction of the ultimate simplicity of nature. The thesis of simplicity involves in addition the assumption that the kinds of small scale elements are few in number… (Bridgman, 1927, p. 51).

For many scientists, pursuing the methodology of analysis or reduction comes with this belief that ultimately nature is simple–the Einsteinian faith that I mentioned at the start.

Experimentally, reflexologists were trying to uncover the simple reflexes by devising special experimental scenarios. One thing that was very important in Sherrington's line of research was to experimentally alter his research animals. Cats and dogs had the cerebral cortex removed, or often he used a preparation in which they were decapitated. The idea

[2] See Pauly (1987) for extended discussion.

was that by removing the highest centres, which were sending inhibitory responses down to the spine, the simple reflexes would be revealed in all of their elementary purity. A picture of an actual device which was used in decerebration operations, from my local museum in Edinburgh is shown in Fig. 1. Loeb used invertebrates, animals with small nervous systems, as prototypes of the simple reflexes. The idea was that invertebrates without all of the higher brain centres which can send inhibitory responses and have all kinds of learning mechanisms, will reveal reflexes in their most elementary form. Sherrington's idea was that when a fully intact dog performed complicated movements, like catching a ball, there was a concatenation of simple reflexes, coordinated to produce the movement. By understanding the simple reflexes, by studying a decerebrated or decapitated dog, the hope was that you'd be able to use those as building blocks for understanding complex motor behaviours.

From Pavlov's version of the theory most of the readers are probably familiar with the contrast between unconditioned and conditioned reflexes. Unconditioned reflexes are the reflexes that an animal is born with, such as salivation with food, which all puppies have. Through the process of learning, the conditioned reflexes are trained in through association between unconditioned reflexes and novel stimuli, like a bell. As I mentioned before, this involved physiological changes in brain pathways, according to Pavlov's theory. But the very possibility of conditioned reflexes rested on this prior level of innate unconditioned reflexes, which were the building blocks for all of the complex learned behaviour that a dog can show.

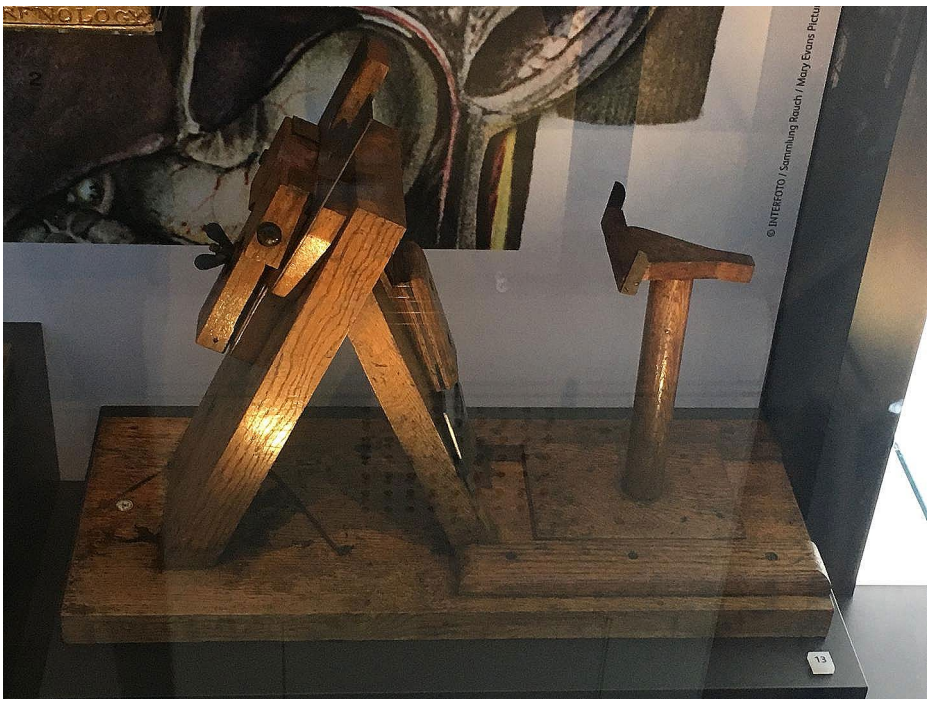

Fig. 1 Sherrington's cat decerebrator 1900's. national museum of Scotland, Edinburgh. Author's own photograph

Criticisms of Reflexology

The reflex theory was dominant in its time, but it also had many critics. It is relevant here to look at what those critics said because they often targeted the role of simplicity and the seeking simplicity in the biological sciences, and considered whether biological objects have *sui generis* properties that are not well understood by importing physics approaches. Kurt Goldstein was a neurologist who wrote extensively, and very critically, about the reflex theory. The phenomenologist Maurice Merleau-Ponty, writing in the 1940s, after the heyday of the theory, was very much inspired by Goldstein's criticism of the mechanistic approaches in biology.[3] Delving into the details about what these different lines of criticism were is out of the scope of this paper, and I will just highlight a very important empirical problem that came up for the reflex theory: the lack of stability of the simple reflexes. Reflexologists were committed to the idea that the elementary reflexes would be the stable, basic responses of the nervous system, and yet empirically they seemed to drift around. Even Sherrington noticed how the receptive field of particular reflexes, the part on the skin from which the reflex could be elicited, would shift or was changed depending on the posture of the animal, or other details of the situation. But in terms of what the theory needed, these elementary components were supposed to be stable and fixed. Later I will say more about the need for stability, where we will see something similar in contemporary neuroscience.

Before proceeding, let us summarise what we said so far. The reflex theory treats simple reflexes as elementary components of the nervous system and behaviour, posits response patterns that are stable and which underlie the apparent complexity of behaviour that is manifestly varying. The physics idea imported there is the method of analysis of classical mechanics and the physicalist approach to biology, which takes organisms to be machine-like. In reflex theory, one indeed thinks of animals as being just like mechanical devices, in that for certain classes of stimuli, if you trigger them one way, they will give you, in a very predictable way, the same reflexive responses. The changeability and spontaneity that other biologists find unique about living systems, is downplayed in the physicalist approach to biology.

Though now it seems quite obviously flawed, reflex theory looked compelling to scientists at the time. Given what has later been found out empirically about the brain, it is of course a gross oversimplification to say that all neural activity can be decomposed into simple reflexes. There are always many historical explanations of why a theory at a particular time comes to dominance. With the reflex theory, however, at least part of the answer of why it was so appealing is because it offered this very simple and elegant approach to understanding the nervous system.

From Cybernetics to the Computational Brain

There are also several historical explanations for why the reflex theory was abruptly dropped in the mid-20th century. One of the factors at play was the invention of digital computers and the availability of new mechanical objects, engineered devices, which neuroscientists could use as model systems and analogues for what they were investigating. I will now

[3] See Goldstein (1934/1939), Dewey (1896) and Merleau-Ponty (1942/1967).

discuss the transition into the computational theory of the brain via the movement known as cybernetics.

Nicholas Rashevsky was a Ukrainian physicist turned theoretical biologist. Before World War II he moved to Chicago and was active in setting up a unit of researchers using mathematical methods in neurophysiology and other parts of biology. He was the mentor of Warren McCulloch and Walter Pitts—discussed below—and very explicit in his agenda of adopting mathematical, physics-inspired approaches into biology:

> The important thing in the mathematical method is to abstract from a very complex group of phenomena its essential features and thereby to simplify the problem. The more complex features are then taken care of gradually, according to the degree of their importance and complexity, as second, third, and higher approximations. True, by abstracting, we lose, so to say, contact with reality; but no harm is done by this *as long as we keep it in mind*. We thus see that the complexity of biological phenomena is rather an argument for the use of mathematical methods than *against* it. In the case of a simple phenomenon we may hope to understand it without the use of mathematics, by simple inspection. But in a complex case we are left hopeless without mathematics. (Rashevsky, 1934, p. 178, emphasis original)

His point was that, given the complexity of biological phenomena, the methods of physics were particularly useful precisely because they indicated a way to approximate, idealise and strip away this manifest complexity. As he says, if you take the first approximation, roughly, to be good enough, you can then add further corrections in your second and third approximations. But because biology is so complicated, we need to rely on these mathematical methods. His paper continues with a long defence of idealisation showing its contribution to past successes of physics. It is interesting to note that Rashevsky's work was not particularly well received by his colleagues who were more knowledgeable in biology (Abraham, 2004). They complained that his models were too unrealistic to be of any use, disconnected from empirical observation and lacking in predictive power.

However, via the work of people like McCulloch and Pitts, this way of treating neural systems has come to be extremely significant in neuroscience. A famous paper by McCulloch and Pitts (1943) introduced the world to a very idealised model of neurons, the ancestor of the point neurons, the dendrite-less, axon-less neurons which are now at the heart of today's artificial neural networks. Instead of thinking of individual neurons as very complicated cells with their weird anatomy and biochemical particularity, these models consider the neuron simply as summing inputs, carrying out a calculation, and generating an output, stripping all of its biological complexity away. In their renowned paper, McCulloch and Pitts showed that by building small networks of these very simplified neurons one can have logic gates and from there build up a digital computing machine. They point to an equivalence between the Turing machine and these neural network systems. Commenting on the impact of neural network models, the mathematician and computer scientist Seymour Papert would write:

> The liberating effect of the mode of thinking characteristic of the McCulloch and Pitts theory can be felt on two levels. … On the local level it eliminates all consideration of the detailed biology of the individual cells from the problem of understanding the

> integrative behaviour of the nervous system. This is done by postulating a hypothetical species of neuron defined entirely by the computation of an output as a logical function of a restricted set of input neurons. (Papert 1965/2016, xxxiii)

The liberation comes, as it were, with having an excuse to forget about the detailed biology of individual cells. The only thing that matters for giving functional explanations of how the brain gives rise to cognition, in this approach, is just this simple calculative process that a neuron does, not all of the other messy biology which is known to be there, but that one feels licensed to relegate to the background.

When examining the role of physics in relation to biology and neuroscience in this era, we also need to consider engineering. What we find is physics meeting engineering and, by that way, meeting biology. This is particularly obvious in the work of cyberneticians like Grey Walter, who were spending their time building toy objects like the "tortoise" or "Machina Speculatrix". This was a little device that had a sensor and could wander around a room. Walter argued that just with these little simple devices he could demonstrate some of the elementary behaviours that living systems with actual nervous systems are capable of. Walter, through his own training and background, did have a foot in the reflex theory of the nervous system, and with the design of these machines he was in fact trying to instantiate some of Pavlov's ideas.[4] In this context, engineering was important in that it gave physics-minded researchers a way to import into their frameworks a notion of purpose or function that, if you take a purely physics approach, really seems quite alien. With biological systems, it is natural to think of them as having goals and some level of agency. But what would be called a teleological approach to living systems has been philosophically problematic in modern science. If, however, you consider physical systems that have been engineered to have certain functions and goals, then you have an analogy for talking about living systems as having functions and goals, but now in terms of a purely physical and artificial device. The role of engineering was, in this sense, to bridge the realms of physics and biology. In relation to this, the neurophysiologist Jerome Lettvin writes:

> Ever since biology became a science at the hands of biochemists it has carefully avoided or renounced the concept of purpose as having any role in the systems observed… Only the observer may have purpose, but nothing observed is to be explained by it. This materialist article of faith has forced any study of process out of science and into the hands of engineers to whom purpose and process are the fundamental concepts in designing and understanding and optimising machines. (Lettvin interviewed in Anderson and Rosenfeld 1998, 13)

An interesting contemporary criticism of McCulloch and Pitts came from none other than John von Neumann:

> What is not demonstrated by the McCulloch and Pitts result is equally important. It does not prove that any circuit you are designing in this manner really occurs in nature. It does not follow that the other functions of the nerve cell which have been dropped from this description are not essential. It does not follow that there is not a

[4] See Walter (1953).

> considerable problem left just in saying what you think is to be described. (von Neumann & Burks, 1966, p. 46)

The point is a simple one: the theoretical elegance of McCulloch and Pitts' design does not by itself demonstrate that that is how the nervous system works, and one should therefore be concerned about ignoring numerous factors that are really important to how the nervous system and brain itself operate. The very things that are left out of McCulloch and Pitts' abstraction could be extremely significant later on.

We can connect this issue with the metaphor of foreground and background in a picture, often used by Kurt Goldstein. Goldstein was a holist about biology—he believed that everything that happens in an organism is relevant to everything else, to some degree or another. He emphasised that when one attempts to understand neurological deficits, one would ideally need to know everything about what goes on in the rest of the nervous system. His point is that, strictly speaking, everything going on in the brain is relevant to any particular process that is showing up as a deficit; and likewise, at the level of the entire physiology of the organism. Because of the large number of relevant factors or phenomena in any biological process, the investigator is tempted to just relegate some of them to the background. Practically, one can only put a small number of them in the foreground. In doing this though, it should not be forgetten that the whole background of complexity that is there might be crucial to a further explanation. The point here is that an abstraction like McCulloch and Pitts' model is restricted to a very tiny number of factors and it deals with them so elegantly that you might be lured into forgetting that the rest is still there in the background.

The provocation here is that it is an open question whether unrealistic idealizations in neuroscience today are as fruitful as those in physics, as Rashevsky claimed they would be. They might be like the physics idealizations which really earn their keep, such as the idealisation of a perfectly elastic collision, or they could be misleading idealizations like the one in the joke about the physicist solving the farmer's problem by positing that the cow is a sphere in a vacuum. Without doubt, artificial neural networks as a technology, as an engineering strategy, have been hugely successful and done tremendous things. On the other hand, as a way of understanding how the brain works, I think it is open to debate whether they are revealing fundamental brain processes or instead missing out too much of what is important.

Present: The Changeable Brain

In this section I would like to say something about the role of dynamical systems theory (DST), another physics approach which has been imported into neuroscience recently, and how this again raises some questions about the difference between biology and physics. A thing about dynamical systems theory is that it explains change in the world in terms of laws of nature, which are themselves changeless. Change in DST is movement through a state space, and that state space is itself fixed. This is something that philosophers of science like Ernst Cassirer noticed more generally about the methodology of physics: that it tends to see its objects in terms of very general fixed laws as opposed to conceiving things in terms of

the particularity and one-offness of events.[5] One can think of dynamical systems as attempts to explain change in nature in terms of the changeless. What is interesting and important about the complexity of the brain is that, as an organ in the body, it is ever changing. Like all biological tissue, it metabolises and turns over. The brain is also plastic: the basis of memory is anatomically and physiologically altering throughout life. There has been recent attention in neuroscience to the ways that some aspects of neurophysiology that were once thought to be relatively fixed, turn out to be less fixed than thought.[6] If this was in many ways not that new to experimentalists, the reason why it appeared so is because theories in neuroscience have tended to downplay the amount of changeability that there is in the brain. This is connected to what I call temporal heterogeneity: that across time the brain is heterogeneous with itself. If one can abstract away from that changeability, one has a very tempting way to simplify the brain, a strategy for ignoring the anatomical and physiological complexity that is there just as a result of the brain changing.

In motor cortex neuroscience, this problem of instability has been really important in the course of theory development. One early theory of motor cortex from Georgopoulos et al. (1986) tried to understand how neural responses mapped on to motor parameters by assuming that the neurons in the motor cortex had a fixed preference for certain directions of movements. This turned out to be empirically problematic because associations between response and direction of movement tended to be fairly unstable even just from trial to trial. Theoretically, scientists were looking for stability in single neuron responses, but empirically they were not obtaining it. That problem has bedevilled theorization of the motor cortex. It is interesting to notice that, as the theories developed, rather than giving up on the idea of an underlying set of fixed responses, scientists have kept on looking for stability, now in terms of population parameters rather than at the single neuron level.

As an example of this, let us consider the approach adopted by Gallego et al. (2017). It is one example of using dynamical systems theory, along with dimensionality reduction through principle components analysis (PCA)—and variants thereof—to search for statistical properties of neuronal populations which show consistency not only from trial to trial, but over different movement tasks which have been performed experimentally. While I take no issue here with the cleverness of the analysis, I regard this as an instance of theorising in which researchers are quite quick to isolate these stable population properties, which they call "neural modes", as the explanatory components for their theory of how neural motor cortex activity relates to movement, as opposed to thinking that anything that is changing in the motor cortex will be explanatory valuable. I believe that, working in the background is that very assumption, continuous from that old approach in reflexology, which just posits from the outset that what is stable and basic in neural responses will be explanatorily most significant.

A more radical approach to changeability would be to ask the question whether living beings are fundamentally historical objects. By this I mean the kind of objects that go through their life course in a changing world and have experienced an unfolding of events which never repeats itself. If you think of an economic system like a financial market, it

[5] "Even where the physicist describes a single event, confined to a definite situation in space and moment in time, he is not concerned with the particular as such, but considers it under the aspect of its repeatability" (Cassirer 1929/1957, 409).

[6] There was not long ago a splash in the Atlantic magazine about new results on representational drift in the sensory cortex. See Yong (June 2021).

never repeats itself precisely, but goes through a historical unfolding in time with some degree of similarity to past patterns. What if living beings are like that? And what if that puts limits on how much can be understood of living beings by importing theoretical systems which are always trying to interpret change and dynamics in terms of some laws or elementary components that are inherently stable and fixed?

This contrast between the natural science of physics and the human sciences, especially history, was put forward by the Southwest school of Neo-Kantians, again, roughly a hundred years ago. Wilhelm Windelband and Heinrich Rickert pointed out that physics gives explanations in terms of fixed laws of nature, whereas human sciences are interested in the particularity, the one-offness of the phenomena of investigation.[7] What if biological objects are in fact not so close to the physics side, and rather have something of that historicity inherent in them? In that case, we would need a quite different framework to grasp that.

Q&A Session, Appendix

Jürgen Jost: Thank you, that was a very interesting historical perspective. I would like to put it perhaps somewhat differently. I mean, as you said, physics is often discovered by simple laws, whereas chemistry, materials, and therefore also biology is intrinsically complicated. Not only because it is historical, but also because it draws upon diversity. Now, however, the neural system is concerned with information processing and information by itself is something that can be formulated in relatively simple terms. If you look at biology, I mean, as I said, biology is typically complicated, but there are certain exceptions: genetic laws, the storage in DNA, is simple and is basically the only biological entity that is so simple precisely because it is concerned with processing and storing and transmitting information. Now, since the neural system is also concerned with these aspects, there is a chance that it can be governed by laws that are simpler than most other laws in biology. In the neural system information is essentially transmitted by spikes which are just binary pulses, and also the neurons as brain cells are the most simplest cells that we have in our bodies. All that leads to the conclusions that information processing might be a simplifying device even in biological systems, and therefore that there can be hopes that new biology is covered by simpler principles than much else of biology.

MC: I wanted to ask you, what do you mean by saying that information is simple? Do you mean information theory?

Jürgen Jost: Yes.

MC: I mean that is a very interesting example of the physics importation because it is clearly taking ideas from thermodynamics. I am reluctant to say that that captures what we need to understand by information in a biological context. It is certainly a useful way of thinking about information transmission and in engineering contexts, but it lacks semantics for one thing. It is a statistical theory, which I do not think captures the richness of what information will be for the living organisms, needing to process information to get around in the environment. And I am also reluctant to conceive that in underlying development you have this one simple molecule DNA, because I think that if you look in detail about how development works there are all kinds of epigenetic and feedback processes, looping back from the developing organism to its DNA.

[7] See "Philosophy of Historical Sciences" in Heis (2018).

Jürgen Jost: I am not claiming that semantics of information and all that is simple, that developmental processes are simple, of course they are complex, but the point is that there's information, the background that might be something that is a simplifying device that makes neurobiology not necessarily very simple, but simpler than other areas of biology. And that you can have some chance to have more general or more principal laws in other fields of biology. Well, of course it would not explain everything.

MC: I would agree that in order to do cognition by itself, if that is what you mean information processing, you need to do a certain amount of abstraction and generalisation. So the brain, in virtue of its cognizing, is going to do some simplification of its inputs. But whether that makes it more intelligible than other complex biological systems, that is what I am a bit more sceptical about.

Thobias Bonhoffer: Thank you very much, that was a very interesting talk. Isn't there an additional complication in everything that you mentioned, namely that everything that we are doing we are also trying to understand with brains again. So you could also rephrase what Einstein said: that he deeply believes that, you know, nature will be simple, that the only thing about nature that we will be able to understand is simple just because our brain cannot grasp infinite complexity. In the sense, if you have deep neural nets for instance, and they learn something, in the end you will not exactly understand what they do because this network is again so complicated that we, with our brains, will not be able to understand it. I think that there is this additional layer of complication that is not how nature actually is, but what we will be able to understand about nature with brains as we have them.

MC: Yes, that is the conclusion that I have ended up with on this: that the commitment to the simplicity of nature has, underlying it, a sort of realisation that if nature is to be fully intelligible to us as finite beings, then it would have to be simple.

Oron Shagrir: Okay, thank you for the very interesting talk. I just want to raise the obvious philosophical objections that we really have no choice, but to use abstraction, idealisation, and simplification methods when we do science. The questions are whether there are different ways to do simplification. Sometimes it takes you in the wrong way, like the example of reflex theory, but it does not show that simplification in general is wrong. Just that this theory went wrong. Perhaps another question is to what extent you can ignore the biological details. If we think about connectionism, or the quotation from Papert you showed, there one completely ignores the biological details; in contrast to more recent computational neuroscience that takes into account certain biological data. There can be a debate here about extent, but you cannot completely discard simplification. Even in the historical approach of course you perform some degree of simplification.

MC: Yes, sure. I would fully agree that any science is going to rely on some simplifying strategies. And I think the interesting question is about which ones you go for and how radical your abstractions are. Goldstein would be an interesting case in point. What he is challenging is the abstractions that are imported directly from physics, where physics has grown up dealing, relatively speaking, with much simpler objects than biology. If you import a very radical abstraction strategy from physics, then you will be underestimating the number of factors and really relevant interactions. Then, if you want to treat someone with a neurological disorder, you are going to be set up on the wrong track. That does not mean to say that he is not in favour of other ways of simplifying, but the point would be that biology needs more of its own research strategies in order to deal with biological complexity, not relying on these importations from physics. As far as philosophers are concerned, I think not

enough attention has been paid to the way that neuroscience today is reliant on all of these simplifying strategies. My criticism of the philosophy community is that there has been a tendency to interpret neuroscience theories at face value, as if this was the whole story about what is going on and not appreciating how much abstraction and idealisation there is.

Wolf Singer: Thank you for this very nice synopsis. I greatly enjoyed it. Biology or living systems have of course constraints. They have to be resilient to disturbances from the outside world. It turns out that one way to obtain resilience and graceful degradation is to increase complexity, increase forces of self-organisation, which requires a high degree of non-linearity and complexity. So, you see, these evolution systems become more and more complex as you go up, and maybe it is not our purpose to provide analytical solutions to these complex systems because as soon as they become very complex, very non-linear, very dynamical, analytical methods that are used in physics do not work anymore, right? There is maybe another level of explanation that we have to reach, rather than trying to analyse these causal relations, which is not possible in the highly non-linear system.

MC: Okay, yes, that is what the provocation I am putting out: What are the other directions to explore? Thank you.

Acknowledgements I would like to thank the organisers of the workshop 'Physics Modelling Thought', Rocco Gaudenzi and Ariel Furstenberg, for their generous invitation to present my research at the MPI and in this volume. Further thanks to go to the members of the audience at the workshop for thoughtful questions and further discussion.

Author Contributions N/A single authorship.

Funding Not applicable – no external funding received.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable.

References

Abraham, T. (2004). Nicolas rashevsky's mathematical biophysics. *Journal of the History of Biology*, *37*, 333–385.

Anderson, J. A., & Edward Rosenfeld (Eds.). (1998). *Talking nets: An oral history of neural networks*. MIT Press.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity* (2nd ed.). MIT Press.

Boge, F. J., Miguel-Ángel Carretero-Sahuquillo, P., Grünke, & King, M. (2023). Introduction: simplicity out of complexity? Physics and the aims of science. *Synthese* 201. https://doi.org/10.1007/s11229-023-04126-1.
Bridgman, P. W. (1927). *The Logic of Modern Physics*. MacMillan.
Cassirer, E. (1929/1957). *The philosophy of symbolic forms, volume 3: The phenomenology of knowledge*. Yale University Press.
Chirimuuta, M. (2021). Reflex theory, cautionary tale: Misleading simplicity in early neuroscience. *Synthese*. https://doi.org/10.1007/s11229-021-03351-w.
Chirimuuta, M. (2024). *The Brain Abstracted: Simplification in the history and philosophy of Neuroscience*. MIT Press.
Dewey, J. (1896). The reflex arc concept in psychology. *The Psychological Review*, *3*(4), 357–370.
Fearing, F. (1930). *Reflex Action: A study in the history of physiological psychology*. Hafner Publishing Company.
Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). *Neural Manifolds for the Control of Movement Neuron* 94:978–984.
Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). *Neuronal Population Coding of Movement Direction Science* 233:1416–1419.
Goldstein, K. (1934/1939). *The organism: A holistic Approach to Biology Derived from Pathological Data in Man*. American Book Company.
Heis, J. (2018). Neo-kantianism. *Stanford Encyclopedia of Philosophy*https://plato.stanford.edu/archives/sum2018/entries/neo-kantianism/.
Loeb, J. (1900). *Comparative physiology of the brain and comparative psychology*. J. P. Putnam's sons.
Loeb, J. (1912). *The mechanistic conception of life*. University of Chicago Press.
McCulloch, W. S., & Walter Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.
Merleau-Ponty, M. (1942/1967). *The Structure of Behaviour*. Translated by Alden L. Fisher. Boston: Beacon Press.
Papert, S. (1965/2016). Introduction. In *Embodiments of Mind*, edited by Warren S. McCulloch. Cambridge, MA: MIT Press.
Pauly, P. (1987). *Controlling Life: Jacques Loeb & the Engineering Ideal in Biology*. Oxford University Press.
Pavlov, I. (1927/1960). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. New York: Dover Publications Inc.
Rashevsky, N. (1934). Foundations of mathematical biophysics. *Philosophy of Science*, *1*(2), 176–196.
von Neumann, J., & Burks, A. W. (1966). *Theory of self-reproducing Automata*. University of Illinois Press.
Walter, W., & Grey (1953). *The living brain*. W. W. Norton.
Yong, E. (2021). Neuroscientists have discovered a phenomenon that they can't explain. *The Atlantic*https://www.theatlantic.com/science/archive/2021/06/the-brain-isnt-supposed-to-change-this-much/619145/.