Philosophy of Science (2022), **89**, 1196–1202
doi:10.1017/psa.2022.35

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

Explaining Neural Transitions through Resource Constraints

Colin Klein

School of Philosophy, The Australian National University, Canberra, Australia
E-mail: colin.klein@anu.edu.au

(Received 20 August 2021; revised 16 January 2022; accepted 21 April 2022; first published online 24 May 2022)

Abstract

One challenge in explaining neural evolution is the formal equivalence of different computational architectures. If a simple architecture suffices, why should more complex neural architectures evolve? The answer must involve the intense competition for resources under which brains operate. I show how recurrent neural networks can be favored when increased complexity allows for more efficient use of existing resources. Although resource constraints alone can drive a change, recurrence shifts the landscape of what is later evolvable. Hence organisms on either side of a transition boundary may have similar cognitive capacities but very different potential for evolving new capacities.

1. Introduction

Smith and Szathmáry (1997) proposed that we could understand the evolutionary history of organisms in part by thinking about a few *major transitions* in evolution. Major transitions included the jump from replicating molecules to cells, the rise of sexual reproduction, and the origin of multicellularity. Each represents a major change in evolvability, that is, in the possible ways that organisms might evolve. Each is also something of a one-way ticket: increased complexity is difficult to wind back once it has become stabilized.

More recently, several authors have argued that we might also understand the evolution of nervous systems as a series of major transitions. Ginsburg and Jablonka (2019) argue that the shift from limited to unlimited associative learning represents something like a major transition, explicitly on the model of Smith and Szathmáry (1997) and their transition from limited to unlimited heredity. Birch et al. (2020) further develop this idea, with unlimited learning as a "transition marker" for a suite of capacities including conscious experiences. Barron, Halina, and Klein (unpublished manuscript) suggest that transitions might be understood in terms of changes in information flow in nervous systems, positing centralization, recurrence, and lamination as key major transitions. McShea and Simpson (2011) note that while changes in information flow are part of Smith and Szathmáry's (1997) discussion, they remain undertheorized and likely more important as organisms become more complex.

Yet, though major transitions represent a powerful tool for thinking about how organisms and the space of evolvability might change over time, considerable questions remain about how and why particular transitions occur. In the case of brains, the question is relatively pressing, as it seems that even very simple brains possess considerable computational power. To put the problem starkly (and hint at the coming answer), a well-known result from computer science suggests that neural networks with a single hidden layer and a nonlinear activation function are universal function approximators (Hornik et al. 1989). Anything we do with a complex, richly structured brain, then, could be done by a simple neural network of appropriate size. Why not just get bigger?

To make sense of a transition, we must make sense of why a transition to a new and more complex form of organization is favored at some time, given that the interesting benefits of evolvability come at some later time. Smith and Szathmáry (1997, 8) note that "we cannot hope to explain these transitions in terms of the ultimate benefits they conferred." The driver of evolution is usually immediate reproductive advantage.[1] In the case of brains, furthermore, it seems that merely making a new function—evolving a new sense organ or new processing trick, or any bread-and-butter evolutionary improvement—does not obviously get one any closer to a major transition. A transition is more than the aggregation of individually useful functions; it is a change in which functions are possible in the first place.

So the fan of major transitions in neural evolution would appear to face something of a mild puzzle. Transitions appear to have occurred and been important, and yet it is not obvious why or how they might occur. Now, this is obviously the sort of thing that should have a solution. As with most evolutionary puzzles, the key is finding something that might generate the necessary fitness gradient. In what follows, I will sketch an argument that a major driver of transitions could be general resource constraints. In section 2, I sketch a brief theory of explanation by resource constraint. In section 3, I then show how a transition might occur using purely resource considerations. I conclude in section 4 with a return to universal approximator theorems and the role of resource thinking in constraining our explanatory approach to neural evolution more generally.

2. Resource explanations

Resource explanations are, broadly speaking, a kind of mechanistic explanation. Mechanistic explanations decompose entities into parts, the coordinated activity of which explains some characteristic activity of the whole (Craver 2007; Bechtel and Richardson 2010). When we break down a mechanism into spatiotemporal parts, however, we find that these often come in one of two different flavors. To explain how an internal combustion engine produces rotary motion, we have to talk about pistons, valves, and injectors, on one hand, and gasoline and air and oil, on the other. These play different roles within the broader explanation and need slightly different

[1] Whether this is true of *all* changes in evolvability is something of a contested question; see Pigliucci (2008) for a review of the conceptual landscape. I follow Pigliucci's conclusion that the sense of evolvability at issue in major transitions is one that can only be selected for indirectly. In this article, I take no particular stance on whether evolvability also requires substantial developmental changes; see Brown (2014) for a helpful discussion of this issue. For what it's worth, I am also assuming that simple saltationist models, which assume a large de novo jump, are off the table.

https://doi.org/10.1017/psa.2022.35 Published online by Cambridge University Press

treatment. Following Klein (2018), I will distinguish between *mechanical parts* and *resources*. Very roughly speaking, this corresponds to an agent–patient distinction: mechanical parts do things (to other parts or to resources), whereas resources have things done to them. We can sharpen up that characterization along four different criteria.

First, mechanical parts tend to *persist* over the timescale of the explanation, whereas resources are often transformed. Gasoline and air are mixed and burned; the pistons and valves stay the same throughout. Note that transformation need not be dramatic or irreversible: cool oil is fed into the engine block and is warmed as it removes heat. It is later cooled, and the cycle continues. Second, mechanical parts tend to be *individual*, whereas resources are often *aggregate*. There are four pistons, and each of them needs to do the right thing at the right time. By contrast, gasoline is a mass that is broken into smaller bits, as needed. Although continuous resources are the cleanest cases, we can have discrete resources as well. A youth soccer team needs so many oranges at half time: the individual identity of the oranges is not important, only that there are enough for everybody. Third, mechanical parts are often *realization indifferent*, whereas resources are often *realization sensitive*. Spark plugs and valves are classic functionalist examples: for the purposes of explanation, nobody cares that much about what they are made of, so long as they are made of something that can do the job. Gasoline, by contrast, is not fungible: put in diesel, and the engine won't work.

Note that the mechanical part–resource distinction—and therefore the satisfaction of the first three criteria—is often explanation relative. If I care how my car stops, brake pads are mechanical parts: persistent, individually important, and functional. If I'm managing a racing team, I may go through so many brake pads that I treat them as a consumable resource. Whole mechanisms in one context can be mere resources in another: the Jeep is a complex whole to the mechanic, matériel to the quartermaster. The point of distinguishing mechanical parts and resources is thus not to draw a firm metaphysical line in the world but to emphasize the different roles that different spatiotemporal parts of a mechanism can play within the same explanation.

Fourth and finally, mechanical parts are usually *causally conservative*, whereas resources are *causally promiscuous*. Each valve has a fixed role and interacts with a few different things in predictable ways. Indeed, an important rule of thumb in engineering is to keep systems modular, if possible, that is, to minimize the ways in which mechanical parts interact (Simon 1996; Calcott 2014). By contrast, resources are often used by multiple different processes at once. In most cars, engine oil both lubricates and cools the engine. The functions can interact: if the oil gets too hot, it ceases to be a good lubricant.

Indeed, in many domains, there is a familiar possibility of *resource competition* with attendant *resource starvation* as an explanation of failures. It takes so much rangeland to support so many head of cattle. Why? Any individual animal would need much less, but the fact that they are all competing for the *same* grass means that starvation threatens on a smaller plot. Conversely, many engineering problems arise from the need for *resource management*. Electrical grids would be simple if there were one producer and one consumer. Delivering energy to many different households with different patterns of demand requires considerable infrastructure just to make sure that everyone gets what she needs.

I have written mostly of concrete resources. However, the point can easily be generalized to more abstract resources. Economics is all about the interaction between various concrete resources (pigs, whiskey, fireworks) and various abstract ones

https://doi.org/10.1017/psa.2022.35 Published online by Cambridge University Press

(money, futures, derivatives). A key set of abstract resources is found in computational theory. In particular, computational complexity theory studies how different algorithms require different amounts of time, memory, processor cycles, bandwidth, and so on (Aaronson 2015). Sometimes these resources trade off against one another: we might cache results from a computation to save time at the expense of space. They also enter into explanations that involve resource competition and resource starvation. My poor choice of a sorting algorithm means that my computer used too much memory, which explains why the operating system crashed. Both would have been fine on their own, but the fact that they were competing for the same, causally promiscuous pool of memory means that they couldn't coexist.

3. Transition and elaboration

Return to the evolution of brains. Brains—and nervous systems more broadly—are incredible resource hogs. Raichle and Mintun (2006, 467) estimate that the human brain uses about 20 percent of a person's total energy expenditure despite being only about 2 percent of body weight. Furthermore, the majority of that is a standing cost: it is paid wheneverone is awake, regardless of what one is doing. In Sterling and Laughlin's (2015) survey of overarching principles of neural design, they show that resource constraints shape even the most basic nervous systems. Some of these constraints also have important nonlinearities. Notably, both the energetic and the volumetric costs of sending information rise disproportionately as the rate increases (Sterling and Laughlin 2015, 54).

The basic question about transitions, recall, was why a change in information flow might be favored, given that the obvious benefits of such a change to the evolvability of new functions don't arise until after the transition. Put in the language of the previous section, transitions may well allow for the evolution of new mechanical parts, but the possibility of new parts cannot be the reason why the transition occurs. An obvious place to look for the answer is instead with resource constraints. Furthermore, as I'll argue, it is possible for a transition favored on resource grounds alone to facilitate new patterns of information flow and thus new functional parts down the line.

Here is an example, using artificial neural networks, of the sort of thing I have in mind. Consider a recurrent neural network containing s neurons and requiring t time steps to calculate some function f. A well-known result shows that this network can be "unrolled" into into a purely feed-foward network, which computes f containing t layers and st neurons.[2] Assume that the timescale is comparable in each case (that is, that recurrent loops take the same amount of time as additional layers), so that the recurrent and unrolled networks also take the same amount of time to compute f.

Now, flip this picture around. Suppose an organism with purely feed-forward connectivity calculates some f by circuit N but that f could be calculated by a recurrent

[2] I use the formulation found in Šíma and Orponen (2003, 2746). Very literal readings of this should be taken with a grain of salt. The result is typically attributed to Savage (1972), with Goldschlager and Parberry (1986, 56) the first to dub it "unrolling." Both the Savage and the Golschlager and Parberry articles concern networks of traditional Boolean gates, however, and these do not need to be trained. The fact that useful trainable recurrent nets use specialized gates, higher-order weight functions, or other similar departures from simple perceptron models suggests the need for more nuance. That said, there are a number of intuitive presentations of recurrent neural networks in terms of unrolling to purely feed-forward neural nets, and the idea that a recurrent net can be translated into a feed-forward net with a space penalty roughly *proportional* to time should be uncontroversial.

https://doi.org/10.1017/psa.2022.35 Published online by Cambridge University Press

network N', such that N is the unrolled version of N'. Ex hypothesi, both circuits calculate the same function in the same amount of time. However, if N has several layers, the shift to N' might come at a substantial energetic savings, because N' would use $s(t-1)$ fewer neurons.

In fact, the trade-off is a bit more complex. Artificial recurrent networks are harder to train: simple back-propagation faces a problem of vanishing or exploding error gradients.[3] Recurrence in biological systems faces an analogous problem, one assumes, because of the inherent instability of excitatory feedback loops. Long-range feedback might also mean more long-range wiring costs, which are themselves a substantial resource drain (Sterling and Laughlin 2015) and one that brains seek to minimize (Cherniak et al. 2004). So additional steps are needed to make a recurrent network stable and trainable. Let's assume that these costs scale with the number of neurons by some constant factor c. Then, we might expect an evolutionary transition from N to N' to be favored, on resource grounds alone, just when $st > sc$.

In other words, we should not expect recurrence to evolve and stabilize when the additional costs of stabilizing a recurrent network are more than the cost of the additional layers needed by a feed-forward network. Furthermore, because these trade-offs are vague and approximate, if recurrence *does* occur, we should expect there to be some point where both N and N' are live options; that is, N' does not have an obvious advantage over N, despite added complexity.

Or, to be more precise, N' does not have an advantage *with respect to computing* f. *However, the transition to a recurrent network might bring benefits for computing* other *functions, for* N′ *can compute functions that take longer than* t *without having to add additional hardware. What* N′ *can do with time,* N *must do with space—and time is often cheaper than space. That is, if an organism has the leisure to run a function for longer, it gets the benefit of recurrence for minimal additional metabolic cost. Adding neurons and wiring, by contrast, adds a substantial fixed cost. Furthermore, the choice of whether to run an algorithm longer can be made on the fly, whereas making a bigger brain usually requires developmental changes.*

For a concrete example of the sort of algorithms that benefit from more time, one might consider what Zilberstein and Russell (1996) call *anytime algorithms*. These approximate a certain function and do a better job the more time they are given, and hence they "allow computation time to be traded for decision quality" (Zilberstein and Russell 1996, 181). Some algorithms of this sort, such as Newton's method for finding roots, are well known. However, there are interruptible anytime versions of algorithms for many problems faced by real-time control, such as the traveling salesman problem (Zilberstein and Russell 1996, 190ff.).

My claim is that a network like N' might get the benefits of these algorithms effectively for free, whereas N has no obvious way to perform additional iterations aside from adding more layers (and thereby committing more resources). That in turn

[3] As Schmidhuber (2015, 93ff.) notes, this problem was known by the late 1980s, received formal expression by Hochreiter's (1991) PhD thesis, and was the focus of intense research for nearly twenty years before recurrent neural networks were competitive at major contests. The advances required to make recurrent neural networks competitive were not simply increases in computational power (though that helped) but also fundamental algorithmic advances. Once recurrent neural networks were feasible, however, they rapidly came to dominate at many tasks—reflecting the argument of this article in miniature.

https://doi.org/10.1017/psa.2022.35 Published online by Cambridge University Press

opens up the possibility of real functional change by making possible algorithms that would be too costly to be useful in simple feed-forward networks.

This is all a how-possibly explanation, of course, and an abstract one at that. The point is not to make claims about an actual transition but rather to show how resource considerations *alone* could support a neural transition. The foregoing explanatory pattern suggests that the transition itself might be favored on pure resource competition grounds. The same function f is computed in the same amount of time before and after the transition. However, the *consequence* of the transition may well be the evolvability of more complex, more efficient, or more useful functions—functions that the original network may resist evolving precisely because of the added cost.

4. Conclusion

In setting up the problem of transitions, I noted the universal approximator theorem and suggested that anything a complex brain does could also be done by a sufficiently large simple brain. Why not just make a simple big brain then? The answer is that big brains are costly, and at some point, the benefits of simplicity are outweighed by those costs. The transition to a more complex pattern of information flow may solve a proximate resource problem. In doing so, however, it may open up the possibility of evolving more sophisticated and more complex functions. A resource-driven transition might change the pattern of evolvability more generally, then, just as the major transitions framework predicts.

I note that though resource pressures are particularly pressing for brains, they have been cited as drivers for more basic transitions as well. Knoll and Hewitt (2011) have an excellent discussion about how many features of multicellularity are driven by the limitations of passive diffusion as a transport mechanism. Once a multicellular organism gets large enough, it cannot rely on simple gradients of nutrients from the outside to the inside. Although there are short-term fixes, a common pathway seems to lead to developmental changes, which ultimately lead to the specialization of function that is a common feature of increasing complexity (Calcott 2011). Again, this strikes me as the sort of transition that is fundamentally driven by pressures on resource management and resource allocation.

Indeed, though I have focused on the efficiencies to be gained by a transition from purely feed-forward to recurrent networks, I suggested that there are costs to be borne as well. In many engineered systems, problems of resource competition are solved by systems dedicated to resource management: if everyone in the house wants to stream a movie at once, a good router will try to balance the load to make sure that no one person saturates the connection. As systems get more complex, more and more effort must be devoted to resource management, including higher-order problems of resource management. Money helps solve the problems arising from the allocation of scarce concrete resources; banks help solve problems that arise from managing large quantities of money; regulators manage banks, and so on. Each level of this hierarchy uses some of the very resources it manages (bankers like to get paid), which in turn creates more complex resource management problems.

We should not expect the brain to be different. The transition to more complex brains comes with increasing pressure on resource management. Indeed, in complex brains like ours, I suspect this becomes a central preoccupation. Resource

https://doi.org/10.1017/psa.2022.35 Published online by Cambridge University Press

explanations of ever-increasing complexity might therefore be the key to understanding major transitions in neural evolution.

Acknowledgments. Thanks to Andrew Barron, Rachael Brown, and Marta Halina for detailed feedback on an earlier draft. This work was supported by a grant from the Templeton World Charity Foundation (TWCF0539). The author declares no competing interests.

References

Aaronson, Scott. 2015. "Why Philosophers Should Care about Computational Complexity." In *Computability: Gödel, Turing, Church, and Beyond*, edited by B. Jack Copeland, Carl J. Posy, and Oron Shagrir, 261–327. Cambridge, MA: MIT Press.

Bechtel, William, and Robert C. Richardson. 2010. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press.

Birch, Jonathan, Simona Ginsburg, and Eva Jablonka. 2020. "Unlimited Associative Learning and the Origins of Consciousness: A Primer and Some Predictions." *Biology and Philosophy* 35 (6):1–23.

Brown, Rachael L. 2014. "What Evolvability Really Is." *British Journal for the Philosophy of Science* 65 (3):549–72.

Calcott, Brett. 2011. "Alternative Patterns of Explanation for Major Transitions." In *The Major Transitions in Evolution Revisited*, edited by Brett Calcott and Kim Sterelny, 35–52. Cambridge, MA: MIT Press.

Calcott, Brett. 2014. "Engineering and Evolvability." *Biology and Philosophy* 29 (3):293–313.

Calcott, Brett, and K. Sterelny, eds. 2011. *The Major Transitions in Evolution Revisited*. Cambridge, MA: MIT Press.

Cherniak, Christopher, Zekeria Mokhtarzada, Raul Rodriguez-Esteban, and Kelly Changizi. 2004. "Global Optimization of Cerebral Cortex Layout." *Proceedings of the National Academy of Sciences of the United States of America* 101 (4):1081–86.

Craver, Carl F. 2007. *Explaining the Brain*. New York: Oxford University Press.

Ginsburg, Simona, and Eva Jablonka. 2019. *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. Cambridge, MA: MIT Press.

Goldschlager, Leslie M., and Ian Parberry. 1986. "On the Construction of Parallel Computers from Various Bases of Boolean Functions." *Theoretical Computer Science* 43:43–58.

Hochreiter, S. 1991. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München. Advisor: J. Schmidhuber.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer Feedforward Networks Are Universal Approximators." *Neural Networks* 2 (5):359–66.

Klein, Colin. 2018. "Mechanisms, Resources, and Background Conditions." *Biology and Philosophy* 33 (36):1–14.

Knoll, Andrew H., and David Hewitt. 2011. "Phylogenetic, Functional, and Geological Perspectives on Complex Multicellularity." In *The Major Transitions in Evolution Revisited*, edited by Brett Calcott and K. Sterelny, 251–70. Cambridge, MA: MIT Press.

McShea, Daniel W., and Carl Simpson. 2011. "The Miscellaneous Transitions in Evolution." In *The Major Transitions in Evolution Revisited*, edited by B. Calcott and K. Sterelny, 19–34. Cambridge, MA: MIT Press.

Pigliucci, Massimo. 2008. "Is Evolvability Evolvable?" *Nature Reviews Genetics* 9 (1):75–82.

Raichle, Marcus E., and Mark A. Mintun. 2006. "Brain Work and Brain Imaging." *Annual Review of Neuroscience* 29:449–76.

Savage, John E. 1972. "Computational Work and Time on Finite Machines." *Journal of the ACM* 19 (4):660–74.

Schmidhuber, Jürgen. 2015. "Deep Learning in Neural Networks: An Overview." *Neural Networks* 61:85–117.

Šíma, Jirí, and Pekka Orponen. 2003. "General-Purpose Computation with Neural Networks: A Survey of Complexity Theoretic Results." *Neural Computation* 15 (12):2727–78.

Simon, Herbert A. 1996. *The Sciences of the Artificial*. 3rd ed. Cambridge, MA: MIT Press.

Smith, John Maynard, and Eörs Szathmáry. 1997. *The Major Transitions in Evolution*. New York: Oxford University Press.

Sterling, Peter, and Simon Laughlin. 2015. *Principles of Neural Design*. Cambridge, MA: MIT Press.

Zilberstein, Shlomo, and Stuart Russell. 1996. "Optimal Composition of Real-Time Systems." *Artificial Intelligence* 82 (1–2):181–213.

Cite this article: Klein, Colin. 2022. "Explaining Neural Transitions through Resource Constraints." *Philosophy of Science* 89 (5):1196–1202. https://doi.org/10.1017/psa.2022.35