A Mathematical Bias

Nicolas P. Rougier, Cédric Brun, Thomas Boraud

A Mathematical Bias

Nicolas P. Rougier[1,2,3,†,*], Cédric Brun[3,4,†] and Thomas Boraud[3,†]

[1]Inria Bordeaux Sud-Ouest — [2]LaBRI CNRS UMR 5289, Université de Bordeaux — [3]Institute of Neurodegenerative Diseases, CNRS UMR 5293, Université de Bordeaux — [4]Université Bordeaux-Montaigne — [†]Equal contributions — [*]Corresponding author: nicolas.rougier@inria.fr.

Abstract. Any Mathematical framework inside which we formalize explanations limits the scope of behaviors we may consider and the type of explanations we might provide. The problem is even more acute when a model provides accurate predictions. This naturally leads us to restrict our investigation within the chosen framework, making us blind to alternatives.

Introduction

In his book "Vision", David Marr [1] proposed *three levels at which any machine carrying out an information-processing task must be understood before one can be said to have understood it completely*. Namely, computational theory, representation and algorithm & hardware implementation. This three-levels analysis has durably influenced the neuroscience community during the last decades and is still the center of intense discussion among researchers, especially concerning the connections between the three levels. We won't discuss the details of the argument since our objective here is rather to address the first recommendation of Marr regarding the goal of computation. We deliberately emphasized *the* because the original phrasing of Marr suggests implicitely two things: (i) there is a goal and (ii) it is unique. If we consider a man-made device observed by another human being, the existence and uniqueness of such device can be certainly acknowledged. However, if we apply the method to a living being, existence and uniqueness are far from being obvious. Moreover, there is a supplementary difficulty when we observe another living being because we're biased by our own cognition and we tend to interpret a behavior in light of our own behavior, and education on that matter. Pareidolia, that is, our tendency for identifying the perception of a stimulus as a face or an object, is a well known trait of human perception. Such incorrect or biased perception has been particularly well illustrated by the experiment of Heider and Simmel [2] using a more complex apparatus. They shot a short movie displaying three geometrical shapes wandering around the screen. With the exception of three subjects in the three groups of observer, all subjects interpreted the movie in terms of actions of animated beings. This is well known in cognitive science and researchers try their best to not interpret animal behavior using their own agenda. And yet, we pretend we may have a different but stronger bias when observing and modelling a behavior. That is, a bias that is not linked to our cognition, but rather to our education and the language we chose for Science, i.e. Mathematics. At this point, maybe a prudent move would be to first re-assess the unreasonable effectiveness of mathematics in the natural sciences [3] before exposing our arguments. Our goal is not to refute the effectiveness nor the usefulness of Mathematics in Science, but rather to explain how their very abstraction power may actually hinder our understanding of brain and cognition.

Taking Advantage of Time

Before developing our point, let us first introduce an unknown system that is able to process some input and produce some output sequence. Let's do three experiments with this system:

Experiment	**Input**		**Output**
Trial 1	2, 3, 1, 4	->	1, 2, 3, 4
Trial 2	A, 7, 4, 6	->	4, 6, 7
Trial 3	9, -1, 2, 5	->	2, 5, 9

Obviously, when presented with some input sequence, this system outputs a different sequence that appears to be made exclusively of elements from the input sequence. We may thus ask what is the goal of such system or said differently, what is the system doing ? If you are familiar with arithmetics (and for the sake of the argument, we'll assume that you are), it is hard not to notice that (1) the input sequence is unordered, (2) the output sequence is made exclusively of positive integers and (3) the output sequence is ordered with increasing values. At this point, it is quite legitimate to assume the unknown system is some kind of sorter of natural number. In fact, if you were to perform another thousand experiments, we can guarantee that the results would only confirm this hypothesis. This means that we can now perfectly predict the output of the initially unknown system. Going back to Marr's three level analysis, we can go down to the algorithmic level in order to find out what kind of sorting algorithm the system is using (e.g. bubble sort, bucket sort, etc.) but, independently on the exact nature of this sorting algorthm, we know that the implementation involves some comparisons (inter or intra) or the moving/swapping of individual values. Having now a full description of the system at all three levels, does that mean we have understood it?

Let us now introduce the real system which is a simple echo state network with a single input channel, a single output channel and no feedback. This network has been trained in a very peculiar fashion and has learned to output the input value but with a delay whose amount (in seconds) corresponds to the value of the input value. Said differently, if you feed X to the system, the output will be X after X seconds have elapsed. Let us play with this model. If we feed the sequence "3, 2, 1" at once (in less than 1 second), this model will output 1 after one second, 2 after two seconds and 3 after three seconds such that the output sequence reads "1, 2, 3". In other words, the output sequence is ordered and you can extrapolate that this works for any numerical sequence even though the answer of the system can take a lot of time depending on the input values.

We can further formalize this neural system by writing an ideal model using a very simple Python program:

```
import time, threading
def process(values):
    def delay(value):
        time.sleep(value/100)
        print(value)
    for value in values:
        t = threading.Thread(target=delay, args=(value,))
        t.start()
process([3,2,1])
```

As you can read (if you know the Python language), there is no comparison, nor moving or swapping of values and no sorting per se. The question is then "what is the system doing?" Is it actually sorting values by taking advantage of time or is it merely outputing values with some delay? This is not a purely rhetorical question because if you think the initially unknown system implements a sort algorithm, then you'll search for some evidence at the implementation level such as comparison or swapping of values. Problem is that if the considered system is complex enough, there are good chances to find some incidental evidences (e.g. correlated activities) that will confort your initial hypothesis while it is false. This is actually the current situation regarding the Bayes theory that may be used to predict behavior but this does not mean there exists a Bayesian implementation inside the brain [4], even though people will regularly report to have found some neural correlates. In our simple example, the hypothesized sorting property derives from the (local) linearity of time and mostly exists in the eye of the (educated and biased) observer. Without further information on the system, we cannot decide if it is actually sorting.

Taking Advantage of Space

Let us now consider another illustrative problem. Suppose you want to study how an animal, when presented with two options A and B, can learn to alternatively choose A then B then A, etc. One typical lab setup to study such alternate decision task is the T-maze environment where the animal is confronted to a left or right turn and can be trained to display an alternate choice behavior. This can be easily formalized using a block world as it is regularly used in the computational literature: The question is then what are the mechanisms that allow an animal (or a model) to alternate decisions between left and right? Given the aforementionned task formalization, you cannot really escape some form of memory inside your model. For such a simple problem, a one bit memory is enough and the simplest solution is probably to negate (logically) this one bit memory each time you reach A or B. When located at X, the model has only to read the value of this one bit memory in order to decide to go to A or B. Going back to neuroscience, can we localize this one bit memory inside the brain and identify when and where the switch occurs? After an animal has been trained, selective brain lesions could be made to various cerebral structures such as to identify the one(s) that disrupt the alternating behavior. But there's a fallacy in such reasoning. The fallacy is to equate information and memory, and more precisely working memory. Consider for example a slightly different setup where the T-Maze is transformed into a closed 8-Maze. Supposed you can only observe the white area while

the animal is evolving along the arrowed line (both in observable and non-observable areas). From the observer point of view, the animal is turning left one time out of two and turning right one time out of two. Said differently, the observer can infer an alternating behavior because of its partial view of the system. The question is: does the animal really implement an explicit

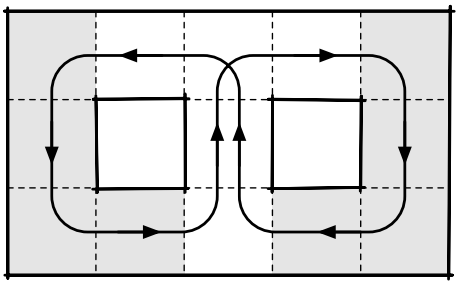

Figure 1: An expanded view of a T-Maze

alternate behavior or is it merely following a mildly complex dynamic path? This is again not a rhetorical question because depending on your hypothesis, you may search for neural correlates that do not exist. This is precisely what happened with working memory that has long been thought to be contained in the sustained activity of neurons in the prefrontal cortex. There are indeed a large number of studies that show such sustained activity and when Rigotti et al. [5] analyzed single neuron activity recorded in the lateral PFC of monkeys performing complex cognitive tasks, they also found several neurons displaying task-related activity. However, once they discarded all the neurons that were displaying task-related activity, they were still able to decode task information using a simple linear decoder. This has been further confirmed by Strock, Hinaut, and Rougier [6] who designed a gated working memory model where the information is encoded in the dynamics of the model such that there is no trace of sustained activity while the information is maintained and can be decoded anytime.

Decision without Value

One aspect of neuroscience that has been deeply impacted by Mathematics is the field of decision making, partly due to its proximity and tight links with neuroeconomics. Probably, also, because of the common mathematical modeling of decision theory that has been borrowed by neuroscientists from the economists. As explained in [7], *nearly all theories of decision, from expected utility theory through prospect theory and even modern reinforcement learning algorithms have shared the notion that in order to choose, the different attributes of each option must at some point be converged, however idiosyncratically, incompletely and imperfectly, into a single value for the actual process of comparison.* This has been further abstracted into the notion of common currency and mapped to a subregion of the ventromedial prefrontal cortex / orbitofrontal cortex. Indeed, using various but finely controlled experimental setup, many studies have shown a correlation between the activity in these regions and the different options present during a choice, making it difficult to argue against such common currency concept [8, 9]. This illustrates quite clearly the influence of Mathematics: decision is equated with value because Mathematics offer no way of comparing

two unrelated objects. Let's take a moment to see what this involves. *Would you rather receive $50 or $100?* This is an easy question and the two options can be easily compared using absolute monetary value. *Would you rather receive $50 for sure or $1000 with 10% chance?* We introduce stochasticity but we can use the expected value instead of the absolute value to make a decision. *Would you rather receive $50 or $100 in one year?* Such temporal aspect can be formalized using the concept of discounted value, where the absolute (or expected) value can be modulated along the temporal dimension and leads to a decision. *Would you rather receive a $50 bill or 10,000 pennies?* Here we have a problem. Absolute, expected or discounted value points to the second option and yet, we can make an educated guess that people would prefer option one. The reason being the different physicality of the two options, one being much more cumbersome than the other. At this point, one could argue that the physicality of an option could be simply integrated in the common currency, provided we find some way of both evaluating and valuing such physicality. But what if we now introduce yet another unforeseen dimension? The story has potentially no end.

Discussion

The mere act of choosing a Mathematical framework inside which we formalize explanations of behavior is not a neutral process. Indeed, mathematical modelling is only an instance of the necessary idealization of the complexity of the brain that is a prerequisite to any scientific approach [10]. In biology, there is a general tendency to project our n dimensional object of interest (were n is unknown) onto an explanatory dimension chosen a priori whereas we should first determine how many dimensions (number of parameters) we need in order to explain our problem and which are the most relevant. Our a priori choice is determined by various biases that are not always consciously addressed such as original background, scientific culture or previous studies but do not rely on a global systemic approach and is therefore doom to fail most of the time because the probability it provides relevant information is inversely proportional to the number of dimensions of the object we study. This indeed implicitly constrains the range of behaviors we may consider and the type of explanations or predictions we might provide. We tend to forget such limits and the problem is even more acute when a model provides accurate predictions or explanations. This naturally leads us to restrict our investigation within the chosen framework in order to explore the limits of a model or an explanation. However, by doing so, we also shadow processes that may completely escape the chosen framework while being nonetheless important if not critical [11]. In that regards, the notion of value that has been proven many times to be fruitful for unraveling some of the inner mechanism of decision making is only relevant to a very specific kind of decision. If we forget this simple fact, we may well provide models or explanations that would cease to be *felicitous falsehoods* [12] and become plainly misleading, by overstepping the perspectival framework in which they have been formulated.

References

[1] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc., 1982. ISBN: 0716715678.

[2] Fritz Heider and Marianne Simmel. "An experimental study of apparent behavior". In: *The American Journal of Psychology* 57 (1944).
[3] Eugene P. Wigner. "The unreasonable effectiveness of mathematics in the natural sciences." In: *Communications on Pure and Applied Mathematics* 13.1 (Feb. 1960).
[4] Matt Jones and Bradley C. Love. "Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition". In: *Behavioral and Brain Sciences* 34.4 (Aug. 2011).
[5] Mattia Rigotti et al. "The importance of mixed selectivity in complex cognitive tasks". In: *Nature* 497.7451 (May 2013).
[6] Anthony Strock, Xavier Hinaut, and Nicolas P. Rougier. "A Robust Model of Gated Working Memory". In: *Neural Computation* 32.1 (Jan. 2020).
[7] Dino J Levy and Paul W Glimcher. "The root of all value: a neural common currency for choice". In: *Current Opinion in Neurobiology* 22.6 (Dec. 2012).
[8] Lotem Elber-Dorozko and Yonatan Loewenstein. "Striatal action-value neurons reconsidered". In: *eLife* 7 (May 2018).
[9] David Spurrett. "Need there be a common currency for decision-making?" In: *South African Journal of Philosophy* 28.2 (Jan. 2009), pp. 210–221.
[10] Olivia Guest and Andrea E. Martin. "How Computational Modeling Can Force Theory Building in Psychological Science". In: *Perspectives on Psychological Science* (Jan. 2021).
[11] Peter Bossaerts and Carsten Murawski. "From behavioural economics to neuroeconomics to decision neuroscience: the ascent of biology in research on human decision making". In: *Current Opinion in Behavioral Sciences* 5 (Oct. 2015), pp. 37–42.
[12] Catherine Z. Elgin. *True Enough*. MIT Press, 2017.