Causal circuit explanations of behavior: Are necessity and sufficiency necessary and sufficient?

Alex Gomez-Marin
Behavior of Organisms Laboratory
Instituto de Neurociencias CSIC-UMH, Alicante, Spain

Abstract

In the current advent of technological innovation allowing for precise neural manipulations and copious data collection, it is hardly questioned that the explanation of behavioral processes is to be chiefly found in neural circuits. Such belief, rooted in the exhausted dualism of cause and effect, is enacted by a methodology that promotes "necessity and sufficiency" claims as the goal-standard in neuroscience, thus instructing young students on what shall reckon as explanation. Here we wish to deconstruct and explicate the difference between what is done, what is said, and what is meant by such causal circuit explanations of behavior. Well-known to most philosophers, yet ignored or at least hardly ever made explicit by neuroscientists, the original grand claim of "understanding the brain" is imperceptibly substituted by the methodologically sophisticated task of empirically establishing counterfactual dependencies. But for the 21st century neuroscientist, after so much pride, this is really an excess of humility. I argue that to upgrade intervention to explanation is prone to logical fallacies, interpretational leaps and carries a weak explanatory force, thus settling and maintaining low standards for intelligibility in neuroscience. To claim that behavior is explained by a "necessary and sufficient" neural circuit is, at best, misleading. In that, my critique (rather than criticism) is indeed mainly negative. Positively, I briefly suggest some available alternatives for conceptual progress, such as adopting circular causality (rather than lineal causality in the flavor of top-down reductionism), searching for principles of behavior (rather than taking an arbitrary definition of behavior and rushing to dissect its "underlying" neural mechanisms), and embracing process philosophy (rather than substance-mechanistic ontologies). Overall, if the goal of neuroscience is to understand the relation between brain and behavior then, in addition to excruciating neural studies (one pillar), we will need a strong theory of behavior (the other pillar) and a solid foundation to establish their relation (the bridge).

«there are empirical methods and conceptual confusions.
Our training and core practices concern research methods;
the discipline is and always has been deeply skeptical of philosophy.
We emphasize methods for the verification of hypotheses and
minimize the analysis of the concepts entailed by the hypotheses. (...)
All the empiricism in the world can't savage a bad idea.»
(**Hogan 2001**)

Part 1. PRELUDE: explaining explanation

A cat is chasing a mouse all over the house. When we observe this, or any other phenomenon, we can be in the presence of it for its own sake — we can know it in its immediacy. If we are curious, however, we are soon compelled to ask what is going on. Our quest for understanding begins. The rational mind seeks an explanation.

But, what does it mean to explain a phenomenon? The canonical approach says that to explain is to find the cause[1]. It is known since Aristotle that the notion of causality has a quadruple structure: understanding why something is what it is requires to identify its material cause (what it is made of, namely, the tangible substrate from which something can take place; i.e. nerve cells, muscle, bone), formal cause (what it is to be, namely, what something can become without contradicting itself; i.e. being a mouse and a cat, a predator and a prey), efficient cause (what produces it, namely, the source of change; i.e. hunger, scent, sight), and final cause (what it is for, namely, that for which it becomes and perfects itself; i.e. the need to escape the mouse, sustain life, the good). Science, in its re-action against the teleology of the 19th century, eschewed the final cause[2]. Since the formal cause appears uninteresting or incomprehensible, and the material cause is taken for granted, this left the 20th century scientist with the main pre-occupation of dissecting efficient causes. In the case of the cat, to find out what brought about chasing behavior.

When understanding becomes the exercise of determining efficient causality, and efficient causality only, one is then compelled to ask: what *produced* the behavior of the cat[3]? A kid would easily reply that the mouse is the cause; another would say that it is because the cat is hungry that it is going after the mouse. For almost all neuroscientists, it is obvious that it is the cat's brain what caused the cat's body to chase the mouse's body with its little mouse brain inside. Our task here is to expose what is meant when one claims that "neural circuit X causes behavior Y", and then determine to what extent this constitutes a satisfactory endpoint to the explanation of behavior itself.

Interestingly, from an evolutionary perspective, one may dare to claim that the converse is true: isn't behavior what produced the brain? (not to mention that bacteria behave, plants behave and robots behave, none with brain cells). In other words, causality in biology goes both ways. It is only because we prefer to concentrate on the short timescales required by our laboratory experiments that we stress proximate physiological causes at the expense of overlooking ultimate ones, including development and evolution[4]. Tinbergen's four questions remind us that to understand behavior one needs

---

[1] More generally, to explain is to substitute fact by abstraction. The notion of causality is a vast topic, and this is not the place to try to give a complete account. It is of interest to briefly note that it has been claimed that «in advanced sciences (...) the word 'cause' never occurs» (**Russell 1913**). Indeed, modern physics has succeeded by finding laws (quantitative invariant relations of variables) rather than causes (chains of antecedent-consequent events).

[2] «Teleology has been discredited chiefly because it was defined to imply a cause subsequent in time to a given effect. When this aspect of teleology was dismissed, however, the associated recognition of the importance of purpose was also unfortunately discarded. Since we consider purposefulness a concept necessary for the understanding of certain modes of behavior we suggest that a teleological study is useful if it avoids problems of causality and concerns itself merely with an investigation of purpose.» (**Rosenblueth, Wiener and Bigelow 1943**)

[3] Causality need not be equated with necessary connection: «It can grant that there are situations in which, given the initial conditions and no interference, only one result will accord with the laws of nature; but it will not see general reason, in advance of discovery, to suppose that any given course of things has been so determined. So it may grant that in many cases difference of issue can rightly convince us of a relevant difference of circumstances; but it will deny that, quite generally, this *must* be so.» (**Anscombe 1971**). Put plainly: "not being determined does not imply not being caused." (**ibid**)

[4] «We suggest that in many cases in biology, the causal link might be bi-directional: A causes B through a fast-acting physiological process, while B causes A through a slowly accumulating evolutionary process. Furthermore, many trained biologists tend to consistently focus at first on the fast-acting direction, and overlook the slower process in the opposite direction. (…) While A is a proximate cause of B, B may have

to consider more than what is just going on here-and-now[5], but to seriously take into account context and history as nested timescales[6].

If to explain is to determine efficient causality, let's examine its common definition in more detail: the cause of an effect is the set of factors that produce, bring about or make the effect happen. The average neuroscientist, then, reckoning the mouse as a stimulus, chasing as a response, and hunger as an internal state, would devote the quest to try to figure out their neural substrate (as we will argue, he or she is not to blame —but still accountable— for the use of a poor behavioral conceptualization and a weak neuro-behavioral nexus). Then, in the effort to spatially localize efficient causality —a textbook example of the so-called mereological fallacy (to ascribe to a part what only applies to the whole)— it is inside the skull of the cat where the real deal is taken to be. And therefore it is there where it ought to be sought.

We glimpse a sophisticated offshoot of behaviorism, which could be called "neuralism". Behaviorism, in its obsession against the mental, insisted that only what was purely observable and measureable as external acts performed by the organism should be worthy of scientific study. When it became fascinatingly possible to start looking at the inside, neurophysiology added a window to study behavior by making the inside nearly comparably as observable as the outside. Because, no doubt, the inside is uncharted fascinating territory, 21st century neuralism surpassed 20th century behaviorism. At the same time, the former carried and extended the most characteristic bias of the latter: the idea that only what is directly observable —now the behavior of the nervous system— is what is relevantly the matter. Thus, neuralism, in its indifference for (if not disdain towards) animal behavior —the very same phenomenon it itself had set at the very core of its agenda (the cat chasing the mouse)— inverted the imbalance: to the degree that one makes sense of what is going on inside, what is going on outside can be rendered as mere contingency. Content ("things held") was once more divorced from context ("things woven"). Differing in what scientists *can* do (now to manipulate and measure the activity inside), both neuralism and behaviorism coincide in what scientists *want* to do (to establish input-output relations, within a "black box", or not). For the most part, reflexology still dominated their thoughts; stimulus and response as the only realities[7]. Inheriting the successful "response function" theory of physics, scientists in the life and

---

prevailed even before A, and may have ultimately affected A. So why is the reasoning of many biologists seemingly more prone to focus at first on the effect acting on the short-term, physiological time scale explanation and not on the processes that take millions of years to manifest themselves? Is it because of the biologists' training?» (**Karmon and Pilpel 2016**)

[5] «Huxley likes to speak of "the three major problems of Biology": that of causation, that of survival value, and that of evolution - to which I should like to add a fourth, that of ontogeny» (**Tinbergen 1963**) or «behavior is part and parcel of the adaptive equipment of animals; that, as such, its short-term causation can be studied in fundamentally the same way as that of other life processes; that its survival value can be studied just as systematically as its causation; that the study of its ontogeny is similar to that of the ontogeny of structure; and that the study of its evolution likewise follows the same lines as that of evolution of form.» (**ibid**)

[6] «It is now time to ask about its phylogenetic origins. Not because an historical explanation could replace the efficient causes of dynamics, but in order to see how these dynamics came to be actualised.» (**Kortmulder 1998** p123)

[7] «precisely on the condition of limiting oneself strictly to the identity or difference of responses in the presence of such and such given stimuli» (**Merleau-Ponty 1942** p183)

social sciences tried it as an ansatz to the study of the behavior of organisms: vary external (or internal) conditions, and relate them to observed changes in output. Rather than stressing one side of the inside-outside dichotomy, the bias of behaviorism is actually better defined as the idea that behavior must be conceived as a lineal input-output process[8]. For neuralism it is the same. Behavior is not regarded as a process of its own right, but as a by-product of neural activity. Then, facts about behavior are deemed as "just phenomena" (and one often hears: "we are interested in mechanisms!"), which amounts to treating behavior as an "epi-phenomenon" of neural mechanisms. Dawkins' reflection on the neurophysiologist's nirvana is worthy of repetition forty years later[9].

Indeed, changes in the brain "go together with" changes in behavior. Despite lacking two-photon microscopy and ignorant about the existence of neurons, a millennial papyrus reports what could perhaps be considered the first observed neural-motor lesion correlation in history: the realization that what is inside the head "has something to do" with behavior[10]. But, why do we take for granted that behavior is understood to the extent that some part of the brain is shown to "cause" it? And, perhaps most importantly, how does that idea determine our scientific agenda?

A great deal of neuroscience has become "circuit cracking". Since its inception, our training as neuroscientists reinforces the idea that given "a behavior", our job essentially consists in pinning down the neurons responsible for it. This is, of course, reasonable. Yet, note that "a behavior" usually implies whatever one wishes to call "a behavior", namely, an unexamined arbitrariness most recently sanctioned by the relative ease at quantifying "what animals do"; any choice of ours is justified as long as we "put a number on behavior" (a critical point we won't have space to fully address here).

Indeed, if neuroscience is understood literally as "neural science" (the science primarily concerned with the properties of neural tissue), one may do without behavior, at least for a while. Yet, if neuroscience's ultimate goal is to explain how nervous activity *enables* or

[8] «Are we not brought back to the classical problems which behaviorism tried to eliminate by leveling behavior to the unique plane of physical causality?» (**ibid** p131)

[9] «If we look far into the future of our science, what will it mean to say we 'understand' the mechanism of behaviour? The obvious answer is what may be called the neurophysiologist's nirvana: the complete wiring diagram of the nervous system of a species, every synapse labelled as excitatory or inhibitory; presumably, also a graph, for each axon, of nerve impulses as a function of time during the course of each behaviour pattern. This ideal is the logical end point of much contemporary neuroanatomical and neurophysiological endeavour, and because we are still in the early stages, the ultimate conclusion does not worry us. But it would not constitute understanding of how behaviour works in any real sense at all. No man could hold such a mass of detail in his head. Real understanding will only come from distillation of general principles at a higher level, to parallel for example the great principles of genetics— particulate inheritance, continuity of germ-line and non-inheritance of acquired characteristics, dominance, linkage, mutation, and so on. Of course neurophysiology has been discovering principles for a long time, the all-or-none nerve impulse, temporal and spatial summation and other synaptic properties, y-efferent servo-control and so on. But it seems possible that at higher levels some important principles may be anticipated from behavioural evidence alone. The major principles of genetics were all inferred from external evidence long before the internal molecular structure of the gene was even seriously thought about.» (**Dawkins 1976**)

[10] «If thou examinest a man having a smash of his skull, under the skin of his head, while there is nothing at all upon it, thou shouldst palpate his wound. Shouldst thou find that there is a swelling protruding on the out side of that smash which is in his skull, while his eye is askew because of it, on the side of him having that injury which is in his skull; (and) he walks shuffling with his sole, on the side of him having that injury which is in his skull.» (**The Edwin Smith Surgical Papyrus 1930**)

supports behavioral processes, then we should be quite preoccupied with how much "filling in the (neural) gaps" per se entails understanding of behavior all.

You may have noticed how the beginning of a standard neuroscience presentation must include "the neural mechanisms of behavior" rhetoric no-matter-what in the first slide (rhetoric, and thus concerned with techniques and skills on how to succeed in the public sphere and advance one's career rather than with the subject matter; persuasion before precision and allure before clarity). Invariably, the speaker must announce a "circuits-of-behavior reduction". The behavioral phenomenon is not only taken for granted, but also deemed as trivial. Dissecting its neural basis is what guides curiosity and drives research (and attracts funding). In fact, by conflating phenomenon with appearance, mechanism appears real while phenomenon becomes epiphenomenal. One should feel fortunate if behavior survives one more slide. If it does, it is usually in the form of an awkward hybrid of anthropomorphic psychological constructs and ad-hoc quantitative indices, more or less automated and refined. From then on, one is free to abandon the phenomenon (aka, the cat chasing the mouse) in order to move into to the real deal asap: the neural circuits.

If the "how" is postulated to be found in neural circuits as the explanatory cause of behavior —and note that it could be sought elsewhere still as a "how", for instance in biomechanics, metabolism, etc— then one is thinking about a particular notion of efficient cause: «the necessary and sufficient condition for the appearance of something», that «at the presence of which the effect follows, and at whose removal the effect disappears»[11].

Part 2. FUGUE: counterfactual dependence

Indeed, upon "certain interventions" at the neural level, "certain changes" take place at the behavioral level. This theoretical paradigm can be methodologically pursued with great efficacy by means of necessity and sufficiency (N&S) tests. The speed, precision and selectivity of current neural interventions have established such decomposition method as the dominant *explicit* experimental procedure and *implicit* conceptual framework to address brain-behavior relations. The N&S approach to explanation is empowered by today's "manipulate and measure" (M&M) doctrine. Both operate under the presupposition that, if one controls the input as well as possible and then records the output as well as possible, any problem is in principle solvable. From this perspective, improved instrumentation and data collection are the essence of progress.

In other words, the explanation of behavior seems to be contained in claims such as "circuit activity X is *necessary* for behavioral process Y" (or, "if X had not happened, then Y would not have happened") in combination, when possible, with claims such as "circuit activity X is *sufficient* for behavioral process Y" (or, "if X were to happen, then Y would also happen").

The N&S approach is legitimate in principle, popular in practice, procedurally simple, methodologically powerful and conceptually straightforward. While its virtues are often celebrated (sometimes hyped), its problematic points are hardly acknowledged, at least in

[11] See Galileo's definition of cause in (**Bunge 2009** p33)

neuroscience forums. Here we try to explicate the difference between what is *done*, what is *said* and what is *meant* when "circuit X explains behavior Y".

Failure of behavioral function upon inhibition of neural function is interpreted as the later being a necessary condition for the former — the circuit is thus claimed to be indispensable, or necessary. Respectively, emulation of behavioral function upon activation of neural function is interpreted as the later being a sufficient condition for the former — the circuit is thus claimed to be enough, or sufficient. If both N&S hold, the circuit is claimed to be *the cause* of the behavior, which is then regarded as pretty much *explained*. But «all the empiricism in the world can't salvage a bad idea» (**Hogan 2001**).

We must ask to what extent N&S reflect facts or their interpretation[12] (note that different interpretations may generate different experiments). In conflict with the above prescription, in reality, co-occurrence of behavioral activity upon neural activation is one thing, and circuit sufficiency is another. Similar concerns must be raised when making necessity claims. Let's see why, and where these leaps of interpretation lie.

The problem with this simplistic model of causality is that it defines necessary and sufficient in order to create a specific effect. It asserts its own formal cause into the process, isolating that phenomenon from any real process. So it defines conditions for its self-validation. When setting up conditions that demonstrate cause and effect relations that we create, we have imitated the principle, but it is still to be seen whether this reveals the actual conditions for the existence of the natural phenomenon. To put it plainly, our causal manipulations do not produce the behavioral effect: they re-produce it in a given context.

Necessary expresses "what is needed", while sufficient expresses "what meets the need" for something to occur. Necessary is what is required, compulsory, indispensable, not susceptive of being waved. Sufficient is what is enough[13], adequate, unwilling to tolerate any more of something. Remaining ambiguous about whether those needs are primarily of the scientist or of the animal whose circuit is to be claimed necessary and sufficient, N&S conditions are harder to understand than to believe.

Our insistence for objectivity can be an excess of pretense for disinterestedness after so much anthropocentrism. In accounting for the behavior of the cat, it is not the cat that is placed at the center of the explanatory effort, but our own activity as humans, with our biases, interests, and habits. We say "circuit X is sufficient" for the cat to behave but what we really mean —and tragically omit— is that "it is sufficient for us to activate circuit X" in order to observe the cat's natural behavior[14]. In other words, we are not concerned with the myriad of processes that nature puts in confluence, but only with the

[12] «statements of necessity and sufficiency are not fundamental truths about neural mechanisms, but rather are interpretations of experimental outcomes» (**DiDomenico and Eaton 1988**)

[13] «Now "sufficient condition" is a term of art whose users may therefore lay down its meaning as they please. So they are in their rights to rule out the query: "May not the sufficient conditions of an event be present, and the event yet not take place?" For "sufficient condition" is so used that if the sufficient conditions for X are there, X occurs. But at the same time, the phrase cozens the understanding into not noticing an assumption. For "sufficient condition" sounds like: "enough". One can ask: "May there not be enough to have made something happen—and yet it not have happened?"» (**Anscombe 1971**)

[14] Causal accounts reflect the notion of liability in court: "the judge decides that an individual is liable to a certain amount for an action he has *caused*".

narrow element that we need to insert in nature in order to emulate its principle[15]. Thus, the sufficient condition belongs more to us than to the cat. Remaining mostly ignorant about how or why the phenomenon occurs, what is brought to the foreground (thus, all that counts in practice) is what is "enough for me to do" so that the cat chases the mouse once more. Similarly, when we say "circuit X is necessary" for the cat to behave, again we imply that it is imperative for nature to have that circuit at work so as to be able to produce the cat's behavior, while all we showed —and all we can really say— is that "it is necessary for us" to remove circuit X so as to be able to block the natural phenomenon. Still remaining in ignorance (now tamed by the great feeling of control that intervention brings), what is brought to the foreground is what is "indispensable for me to do" so that the cat doesn't chase the mouse this time. Causal explanations of this flavor, then, turn out to be more necessary (and sufficient) for the neuroscientist than for neuroscience; circuit activation and inhibition are more the means for the neuroscientist to try to explain the cat's behavior than for the cat to try to catch the mouse[16].

The sufficient condition *is by itself not sufficient* because the effect it is shown to produce depends on the conditions in which the cause takes place[17]. This is, each of the elements of the sufficient condition may be necessary to prove emulation, while none of them alone is strictly sufficient. That condition, being all that may be required for the scientist to elicit behavior, is not all what is required for behavior to occur. The sufficient condition is then "relatively sufficient". Moreover, since necessity in biology is nearly always relative, the necessary condition is "contingent necessary"[18]. It would then be more accurate to qualify N&S as *relative-contingent* N&S.

Circuit N&S do not occur in the void, but in (and "because of") the particular conditions and relations to other neural and non-neural elements and processes (genetics, biomechanics). One would wish but in fact cannot leave out the contingency of the experimental design (the task that is chosen by the experimenter), nor the temporal

---

[15] «the lights went on when Mrs. Smith turned the switch (…). It is evident that the statements are all singular, rather than general or lawlike. Moreover, in none of them is the occurrence tacitly assumed to be "the cause" a sufficient condition for the event alleged to be its effect. For example, turning a switch does not suffice to produce illumination, since many other conditions must be satisfied for this to happen. In making such causal statements, it is, of course, possible that we know what these further conditions are, and take them for granted without mentioning them explicitly; but this is rarely the case, and we are usually able to cite only a few of these conditions, without knowing all of them. In either case, what we are doing is designating as the cause of an event just one item, selected from what is tacitly supposed to be its full complement of necessary and sufficient conditions, because the item is deemed important for various reasons.» (**Nagel 1965**)

[16] An interesting caveat —revealing an arbitrary preference for a particular level of organization and working in conjunction with a reduced notion of causality— is the following: if circuit activation is said to produce the behavior of the animal, it could also be said that the behavior of the scientist has produced the animal's circuit activation in the first place, or «[a]t most it would show that experience could be produced by means of interaction between a probing scientist and a healthy animal. We haven't yet imagined a case in which experience emerges from the causal effects of neural activity alone» (**Noë 2004** p211)

[17] «other factors, without which the event would not occur, are all assumed to be constant and given.» (**Eaton and DiDomenico 1985**)

[18] «although in the sense of "cause" under discussion the cause of an event is a necessary (or indispensable) condition for its occurrence, the necessity may be only relative. (…) there may be more than one set of sufficient conditions for an event's occurrence (….). Accordingly, in the sense of "cause" under discussion, the cause of an event is in general neither a sufficient nor an absolutely necessary condition for the event's occurrence. The cause may be called a "contingently necessary" condition» (**Nagel 1965**)

constraints of the test (whether manipulations allow time for the organism to adapt, learn or recover). The immediacy of the laboratory conditions in which we test reality often involves shrinking the spatiotemporal spread of behavior[19]. We are often ignorant of (and, unfortunately, sometimes indifferent to) the actual conditions for the occurrence of the behavioral phenomenon we are trying to explain[20]. We can't specify either all the circumstances for which the so-called cause will not cause its effect anymore[21]. "X sufficient for Y" suggests but does not mean "sufficient in a vacuum", yet it implies "everything else being equal". Not knowing the context, what we find in one may not apply in another. Disregard contingency, still this view of causation is dependent on it.

Behavioral absence upon circuit inhibition, and its presence upon activation, can generate false positives and false negatives, and produce spurious conclusions (see **Eaton and DiDomenico 1985**). Results can have alternative interpretations due to the system's plasticity and redundancy. A neuron or circuit may be N&S for the chain of processes to continue, but not for the end result, due to shared responsibility and multiple realizability. Activation and inactivation can have secondary side effects. Moreover, perturbations can be opposed and compensated by the organism.

Another limitation of the N&S approach is that most of its claims are not univocal, but statistical. They are valid only upon averaging of trials and pooling of individuals. Being true for everybody, they are not true for everyone. They are the «confused result of many instances» (**Whitehead 1933** p112). Even accepting the power of M&M under the framework of N&S tests, most of the time we can't "bring about Y by doing X". Variability is then conflated with noise and deemed as undesirable both for the scientist and for the animal. Yet, when control is studied from the perspective of the animal, behavioral variability may be in service of constancy[22].

---

[19] «There is something in the context of the experiment which goes beyond the stimuli and responses directly found within it. There is, for example, the problem which the experimenter has set and his deliberate arrangement of apparatus and selection of conditions with a view to disclosure of facts that bear upon it. There is also an intent on the part of the subject. Now I am not making this reference to "problem", "selective arrangement" and "intent" or purpose in order to drag in by the heels something mental over and beyond the behavior. The object is rather to call attention to a definite characteristic of behavior, namely, that it is not exhausted in the immediate stimuli-response features of the experimentation.» (**Dewey 1930**)

[20] «The point I want to stress is that we seldom have enough information to state explicitly the full set of sufficient conditions for the occurrence of concrete events. The most we can hope to accomplish in such situations is to state what are the best only "important" indispensable conditions, such that if they are realized the occurrence of the designated events is made "probable"; and we thereby take for granted that the remaining conditions essential for the occurrence of the events are also realized, even when we do not know what those remaining conditions are.» (**ibid**)

[21] «They realize that if you take a case of cause and effect, and relevantly describe the cause A and the effect B, and then construct a universal proposition, "Always, given A, a B follows", you usually won't get anything true. You have got to describe the absence of circumstances in which A would not cause a B. But the task of excluding all such circumstances can't be carried out.» (**Anscombe 1971**)

[22] «Instead of asking how a particular experimental manipulation alters the subsequent behavior of an organism, one might instead ask how an experimental manipulation alters the parameters of the system. This is a subtly different question, but the difference is important, and requires that the parameters of the system be understood to begin with. Understanding what variables organisms may be controlling necessitates that organisms be understood on their own terms before they are used as model systems to answer larger questions.» (**Bell 2014**)

The popularity of the N&S approach in neuroscience is in part fed back by the amounts of neural and behavioral data one can now produce at ease. However, the availability of unlimited data, even within putative infinitely precise M&Ms, will not suffice to understand the brain, behavior and their relation. The humbling effort of applying that logic to the functioning of a simple microprocessor cast huge concerns about the validity of such expectation[23], calling into scrutiny the habit of collecting facts *for the sake of it*, in case we may need them in the future[24]. Given the scarcity of brain-behavior theories, we don't know yet what the right levels of granularity and abstraction are. Without a conceptual thrust to explain behavior, it is unlikely we shall be able to "understand the brain". Call it behavioral chauvinism if you wish; the fact is that brains are for behavior.

Notwithstanding the value in manipulation, when our interest is not so much to understand the world but to control it, we end up learning what we do to things rather than what things do. Overall, we don't understand what or why behavior is, but a particular way of inducing its occurrence in hopefully similar experimental conditions. Things get much more delicate when such "things" are animals, because animals are living organisms and their behavior is, as we will see later, self-caused[25]. Actually, what organisms do is to control their perceptions by means of opposing (our and other) disturbances (see below and **Powers 1973**). In the lab, our concern with what *we* can make animals do easily precludes us from knowing what *they* do and why they do it.

But for many, to explain is to successfully intervene: "A causes B if control of A renders B controllable". If I can outsource the happenings of B to my intervention on A, I will then expect a change in A to be followed by a change in B. Similarly, for many others, to understand is taken as to be able to fix what one tries to understand[26], which is an inversion of Stuart Mill's famous prescription: "to find out how something works, look to see how it fails". Building, though, is more challenging than intervening or fixing: to build a model *that behaves* is much more insightful than to make a model *of behavior*. In other words, to "abstract" a behavioral process (via classification) or to "extrapolate" from one case to many (via statistics) is less powerful than to have a generative model[27]. But from the interventionist point of view, explanation is that practical information

[23] «We argue that the analysis of this simple system implies that we should be far more humble at interpreting results from neural data analysis. It also suggests that the availability of unlimited data, as we have for the processor, is in no way sufficient to allow a real understanding of the brain.» (**Jonas and Kording 2016**)

[24] «the rule "collect truth for truth's sake" may be justified when the truth is unchanging; but when the system is not completely isolated from its surroundings, and is undergoing secular changes, the collection of truth is futile, for it will not keep. » (**Ashby 1958**)

[25] «Behaviour, as a relation between a living system operating as a whole and the medium operating as an independent entity, does not take place in the anatomy/physiological domain of the organism, but depends on it. (...) the behavior that a living system exhibits is neither determined by it nor by the medium alone, even when a particular structural change in a living system may specifically interfere with its ability to generate a particular behavior » (**Maturana 1995**)

[26] «To understand what this flaw is, I decided to follow the advice of my high school mathematics teacher, who recommended testing an approach by applying it to a problem that has a known solution. (...) I started to contemplate how biologists would determine why my radio does not work and how they would attempt to repair it. » (**Lazebnik 2002**)

[27] «An explanation is the proposition of a generative mechanism or process which, if allowed to operate, gives rise, as a result of its operation, to the experience to be explained» (**Maturana 1995**)

relevant to manipulation, which does not necessarily imply generalization for prediction, and which is quite different than grasping the concrescence of behavior. Be that as it may, to intervene *is not* to explain[28].

The natural urge to upgrade correlation to causation (a still surprisingly common flaw) has gone further: it has replaced, when possible, neural correlates (NCs) with neural counterfactuals (NCFs), with the concomitant urgency to upgrade intervention to explanation. This makes us believe that what caused the comportment of the organism is our action upon it, by means of altering certain conditions that trigger this or that output. We fail to understand the behavior is control *of* the organism *by* the organism. Such abuse and misuse of causal thinking is susceptible in and typical of a kind of industrialized science, where the balance between cutting-edge instrumentation and intellectual depth is tilted. When convenience is king, we may be tempted to sacrifice understanding[29].

The M&M approach, under the N&S paradigm, has more important and far-reaching provisos. It conflates experimental results with the particular interpretation it makes of them[30]. Running with virtually no theory (something to be covered in more detail in another occasion), it appears to exempt one from making explicit any pre-suppositions for the explanation of the phenomenon; no abstract idea beyond the minimal conjecture that "a change" between two conditions might be observed (and if it is not observed, one can always vary the conditions until it is). Such a "nought of question" easily dissimulates itself under the answers provided by statistical hypothesis testing: testing for rejection of the null hypotheses is transformed into the rejection of testing a null hypothesis (a camouflaging that is not the statistician's fault). Lack of conceptual insight often collapses into a "Shakespearean two-alternative forced choice": to be or not to be... statistically significant.

Now it is clear why the interventionist method at the circuit level doesn't teach us much about the structure of behavior, but mainly whether it occurs or not (regardless of the its biological significance), which confronts us with a grave problem: anything can be called behavior[31]. In other words, when searching for NCFs, "anything goes" for behavior. Erecting the neural circuit causation approach as the conceptual framework to understand behavior reduces the exploration of behavioral theories to the exploitation of

---

[28] «If a biologist or another scientists is interested in control or manipulation, and many of them are, then this is an epistemic goal for that biologist doing science. It is an important goal, and it is important to learn how to intervene and manipulate in experimental settings. Much scientific training and subsequent practice involves pursuing that goal. Having these as goals for one's science and scientific practice is not the same as finding a mechanism nor is it the same as explaining things by mechanisms. In fact, controlling is not explaining at all.» (**Machamer 2004**)

[29] «Its first characteristic is that its ultimate aim is not understanding but the purely practical one of control. If a system is too complex to be understood, it may nevertheless still be controllable. For to achieve this, all that the controller wants to find is some action that gives an acceptable result; he is concerned only with what happens, not with why it happens. Often, no matter how complex the system, what the controller wants is comparatively simple: has the patient recovered? —have the profits gone up or down? —has the number of strikes gone up or down?» (**Ashby 1958**)

[30] «This operational approach excludes alternative methodologies because the N&S causal definition is the same as the methodology for its own demonstration.» (**DiDomenico and Eaton 1988**)

[31] «the methodological connection to behavior is undefined since the Command Neuron Experiment allows virtually any phenomenon to be used as "behavior"» (**ibid**)

handling protocols. The search for principles of behavior is substituted by the development of molecular and cellular techniques. Empirical work takes completely over theoretical work. Concepts are reduced to methods and experimentation becomes data collection. Then, explaining behavior consists in coarsely describing it, saying "because", and then describing in detail some of the molecular and cellular substrates that "produce" it. If we agree to call behavior "anything I can put a number on", it is then likely that we will discover many neural substrates of foggy constructs[32].

A deeper source of confusion and misunderstanding is the belief that the N&S approach belongs to the practice of a kind of "immaculate perception", devoid of and proudly untainted by any bias of interpretation. Being nowadays most popular and least consciously practiced "epsilon of theory" —self-limited to what the eye can literally see, not what the mind can think— it preaches the M&M gospel of our times: "when we map it all, we shall explain it all". But, such lack of premise is nothing but a premise of lack[33].

Pay attention to the Q&A after a seminar, and you may observe the ease with which the speaker happily turns down speculation when asked "what would you expect?" by uttering "we don't know" and then comfortably completing the by saying "we will see what happens when we do this and that". As if not having any idea of what could be going on, and leaving it all to future M&Ms, was a virtue. Their null model is then merely "a difference" in its minimal conceptual sense, namely, that upon changes in one variable we expect something different in another. When, upon scratching, the *null* hypothesis is hardly more than "to check the effects of X on Y", we are in the presence of the *dull* hypothesis: put a thousand cats in a thousand boxes with a thousand mice, and randomly turn a thousand circuits off and on, until behavior disappears and appears again. Lack of insight, both before and after the experiment, is compensated by figures that seek to impress (ie. dozens of automated assays in parallel, hundreds of neurons recorded simultaneously, thousands of animals tested, millions of frames generated, billions of dollars spent). This is a naïve understanding of the idea that "more is better"[34].

Of course one can't in principle be opposed to data, more data, and much more data. In practice, though, this can lead to "chaos in the brickyard" (**Forscher 1963**). Scientists can be considered builders of explanations by means of assembling facts. Such "bricks", difficult and expensive to make, were essential for the edifice not to collapse. As more emphasis was put into the brick making system, data became easier and faster to produce. Very good bricks were certainly made. And many more were produced ahead of demand, pending the decision of what type of building one would be supposed to erect with them. Ultimately, there were so many bricks everywhere that they started to defeat

[32] A biologist friend of mine once told me: «what I care about is to pin down the genes of a behavior, behavior being anything I can put a number so that it will allow me to get to my genes». Replace genes (or gene networks) with neurons (or neural circuits) and the punch-line is the same: behavior as a pretext for genetic and neural manipulations in the age of technocratic science.

[33] «The scientist cannot make the rejoinder here that he thinks without ontological background. To believe that one is not doing metaphysics or to want to abstain from doing it is always to imply an ontology, but an unexamined one —just as governments run by "technicians" do not make political policy, but never fail to have one— and often the worst of all.» (**Waelhens 1942**)

[34] «this is no doubt due to practical preoccupations and notably to a representation of productivity and labor in terms of scale. The more masons work, the higher the building rises. The more copyists copy, the longer their copy becomes. Fabricating labor is the only labor in which the amplification of the product is quantitatively, spatially proportional to the progress of action. (...) A technician's and artisan's thought willingly concentrates on this demiurgic elaboration of the indeterminate.» (**Jankélévitch 2015** p168)

their purpose. «It became difficult to complete a useful edifice because, as soon as the foundations were discernible, they were buried under an avalanche of random bricks. And, saddest of all, sometimes no effort was made even to maintain the distinction between a pile of bricks and a true edifice» (**Forscher 1963**). Data, for data's sake, overrules theory and, thus, nullifies itself.

This radical belief in induction[35], intertwined with a conscious eagerness of not pursuing hypothesis[36], misses the fact that any description starts with comparison which, in turn, assumes a particular perspective[37]. Indifference is impossible. There is always something in the data which we turn our attention to and away from. Accordingly, bringing forth the contributions of our subjectivity is an honest attitude, prone to less confusion than insisting on unbiased approaches (often confusing unsupervised with unbiased) and hoping to be able to produce results that equate with their interpretation. Pretense of absolute is an absolute pretense. And dismissing that as "just philosophy" is just a philosophy of dismissal[38].

---

[35] The belief that «[s]cientific discovery, or the formulation of scientific theory, starts in with the untarnished and unembroidered evidence of the senses. It starts with simple observation —simple, unbiased, unprejudiced, naive, or innocent observation—and out of this sensory evidence, embodied in the form of simple propositions or declarations of fact, generalizations will grow up and take shape, almost as if some process of crystalization or condensation were taking place. Out of a disorderly array of facts, an orderly theory, an orderly general statement, will somehow emerge» (**Medawar 1978**)

[36] «The belief underlying Mass Observation was apparently this: that if one could only record and set down the actual raw facts about what people do and what people say in pubs, in trains, when they make love to each other, when they are playing games, and so on, then somehow, from this wealth of information, a great generalization would inevitably emerge. Well, in point of fact, nothing important emerged from this approach. (...) [T]he starting point of induction is philosophic fiction. There is no such thing as unprejudiced observation. Every act of observation we make is biased. What we see or otherwise sense is a function of what we have seen or sensed in the past. The second point is this: Scientific discovery or the formulation of the scientific idea on the one hand, and demonstration or proof on the other hand, are two entirely different notions.» (**ibid**)

[37] «In the use of language, for instance, we depend on the fact that names have been given to objects, qualities, and relations, which fix certain similarities and differences in the flow of experience as boundaries containing it, dividing it, directing it. Whenever we describe, we class things or properties or events together or apart on the basis of the similarities and differences marked by the words we choose. Consequently, to the extent that science begins with description, it begins with comparison. But no two things, no two qualities, no two events are alike in all respects, or alike in none. Any description singles out some similarities and differences to the exclusion of others, which could be the basis of alternative descriptions. Consequently, a demand for a complete description of anything amounts to a contradiction in terms. A demand for a pure description would be equally incoherent, for, of necessity, the similarities and differences that we pick out when we describe anything will depend on what we intend the description for, our expectations about the matter in question, considerations of relevance to some focus of interest, and other prior assumptions. Comparison necessarily assumes perspective.» (**Beer 1980**)

[38] «While often explicitly denying the relevance of philosophy to its operations, psychology has implicitly used the philosophical assumptions of a seventeenth-century ontological dualism, a nineteenth-century epistemological empiricism, and an early twentieth-century neopositivism, to build a standard orthodox approach to the resolution of the antinomies. (…) the product of the acceptance of some basic ontological and epistemological —hence philosophical—assumptions. These assumptions begin with the idea of splitting reason from observation, and follow with the epistemological notion that knowledge and, indeed, reason itself originates in observation and only observation. These assumptions then lead to a particular definition of scientific method as entailing observation, causation, and induction-deduction, and only observation, causation, and induction-deduction. Sometimes, the split is found in explicit and implicit attacks on theory, as in a particular rhetoric that states that all theories must be induced directly from observations (i.e., must be "data based" or "data driven"). It is also found in a dogmatic retort given to any reflective critique —"that's just philosophy."» (**Overton 2006**)

If what I am saying is accurate, it will also be unpopular, because most biological sciences, and in particular neuroscience, have great investments in counterfactual dependence as a proxy for explanation. The step-by-step interventionist approach is comforting to the intellect because it provides a monotonic succession of measurable chains. The M&M modus operandi and its recipe are this: try to keep everything fixed, change one thing at a time[39], and see what varies. But, as we suggested above and will see below, exploiting the relationship between an independent variable and a dependent variable presupposes a lineal notion of causation, which is inadequate in the study of animal behavior[40]. In biology (the study of living organisms!) the method of investigation must respect the system being studied. Organisms cannot be properly studied as closed physical systems. William James example opposing magnets and lovers[41] emphasizes that one of the essential properties of living organisms is that they can produce the same ends *with* (and *by*) different means. Inert systems, in contrast, tend to produce different ends even with very similar means. The former are characterized by convergence to a goal, the latter by sensitivity to initial conditions[42]. This leads us into a serious appraisal of what *the behavior of organisms* is, and what it is not; the profound difference between a cat chasing a mouse and a leaf blown by the wind.

Part 3. ALLEGRO: beyond lineal causality

In our era of non-risky science due to scientific careers at risk, little can change if what is rewarded (and thus selected for) is confined to the conceptual template: "I study the role of brain region X in the behavioral construct Y" — even if we add some glam bio-tech and really good selling skills... Such a research program provides "many results" which, upon close inspection, are often *questionless answers*. A lot is done but nothing is really tested, and so one can never be wrong. Note how many titles and abstracts in neuroscience journals and seminars reflect (at the same time that conceal) this by means of "filler verbs" that project a sense of understanding and explanation[43]. This lack of

---

39 «until about 1925, the rule "vary only one factor at a time" was regarded as the very touchstone of the scientific method.» (**Ashby 1958**)

40 «What we have is a circuit, not an arc or broken segment of a circle. This circuit is more truly termed organic than reflex, because the motor response determines the stimulus, just as truly as sensory stimulus determines the movement.» (**Dewey 1896** p363)

41 «If now we pass from such actions as these to those of living things, we notice a striking difference. Romeo wants Juliet as filings want a magnet; and if no obstacles intervene he moves toward her by as straight a line as they. But Romeo and Juliet, if a wall be built between them, do not remain idiotically pressing their faces against its opposite sides like the magnet and the filings with the [obstructing] card. Romeo soon finds a circuitous way, by scaling the wall or otherwise, of touching Juliet's lips directly. With the filings the path is fixed; whether it reaches the end depends on accidents. With the lover it is the end which is fixed; the path may be modified indefinitely.» (**James 1890** p6)

42 «A profound difference between most inanimate and living systems can be expressed by the concept of equifinality» (**Bertalanffy 1950**)

43 Amongst the most used verbs we find: "involves, reflects, reveals, mediates, is associated with, contributes to, shapes, modulates, alters, regulates, drives, determines, generates, produces, encodes, underlies, induces, enables, ensures, supports, promotes, suppresses, inhibits, prevents, disrupts, controls, and causes". The pet expression is perhaps "X plays a role in Y", or "The role of X in Y". Certainly, lack of coffee "plays a role" in my writing of this piece, and gravity "mediates" to the mouse escape from the paws of the cat.

(ability or interest for?) a solid theory of behavior and of brain-behavior relations is turning the "explanatory gap" between circuits and behavior into an "explanatory jump". Since neuronal circuits are nominated to *explain behavior*, we were compelled to ask two apparently unnecessary questions: What does it mean to *explain*? And, what is *behavior*? All the effort made in this essay went in the direction of not taking for granted that turning neurons on-and-off and subsequently observing behaviors on-and-off is a satisfying schema for explanation in current neuroscience. Let me conclude by sketching, very briefly, some ideas and alternatives[44] on N&S and M&Ms, and their associated ideologies.

An organism is cause and effect of itself — one can say it louder but not clearer. Behavior is control. Behavior is, therefore, a circular-causal process. Then, the study of living beings requires going beyond lineal causality[45]. The foundations to understand feedback control can actually be found way back in the history of Egyptian inventions[46]. Explanations of behavior need to abandon the "push-pull" idea of lineal causality and embrace the "go-round" notion of circular causality (ironically, the etymology of circuit is to "go-round").

If we must cease to look for cause and effect in behavior, what shall we look for then? When it comes to studying the behavior of organisms, lineal causality is to circular causality what arithmetics is to algebra. Namely, rather than being engaged in finding all possible combinations between stimulus and response, the real quest is to find stable relationships, whatever the value of the variables[47]; to find relations, rather than list all connections. Data collection then becomes experimentation as well as conceptual testing.

---

[44] "But it is not sufficient to oppose a description to reductive explanations since the latter could always challenge these descriptive characteristics of human action as being only apparent. It would be necessary to bring to light the abuse of causal thinking in explanatory theories and at the same time to show positively how the physiological and sociological dependencies which they rightly take into account ought to be conceived." (**Merleau-Ponty 1942** p176)

[45] «In describing the physical or organic individual and its milieu, we have been led to accept the fact that their relations were not mechanical, but dialectical. A mechanical action, whether the word is taken in a restricted or looser sense, is one in which the cause and the effect are decomposable into real elements which have a one-to-one correspondence. In elementary actions, the dependence is unidirectional; the cause is the necessary and sufficient condition of the effect considered in its existence and its nature; and, even when one speaks of reciprocal action between two terms, it can be reduced to a series of unidirectional determinations. On the contrary, as we have seen, physical stimuli act upon the organism only by eliciting a global response which will vary qualitatively when the stimuli vary quantitatively; which respect to the organism they play the role of occasions rather than of cause; the reaction depends on their vital significance rather than on the material properties of the stimuli. Hence, between the variables upon which conduct actually depends and this conduct itself there appears a relation of meaning, an intrinsic relation. Once cannot assign a moment in which the world acts on the organisms, since the very effect of this "action" expresses the internal law of the organism.» (**Merleau-Ponty 1942** p160)

[46] «Ktesibios's water clock required a steady, unvarying flow of water to measure accurately the steady, unvarying flow of time. But because water flows more quickly from a full container and more slowly when it is less full, Ktesibios had to devise a way to keep the vessel at a constant level while water was flowing from it into the clock mechanism. As he did this in a manner not unlike that of the modern flush toilet to which it is assumed the reader has handy access, I will use this more modern invention instead of the water clock as our first example of a feedback-control device.» (**Cziko 2000**)

[47] «When you learn to see behavior in terms of relationships among variables instead of causal connections between one event and another, you can see invariance where formerly you could see only specific causal connections. You can see that when someone builds a fire in the fireplace, two people may show "the same behavior", even though one of them takes off a sweater while another opens a window.» (**Powers 1998**)

Accordingly, to treat behavior as a dependent variable, and stimulation or neural manipulation as an independent variable, ignores feedback, which is the essential ingredient in the process that constitutes behavior. The M&Ms approach (precise control manipulates one variable in the system, and precise monitoring captures whatever the change), which by itself cannot suffice, is still valuable if N&S tests are upgraded to the Test for Controlled Variables (**Marken 2001**). The ability of the organism to oppose disturbances, granted by circular causality via negative feedback loops, allows to revisit the taboo of teleology in the science of behaving systems[48]. Rather than concentrating on models of behavior, let us conceive models that behave, and then set up our experiments so as to allow control, not of the animal, but by the animal. In a word, to see behavior from the animal's perspective[49].

Another way to phrase this substantially different alternative is the following: adapt the method of study to the object of study (rather than the current Procrustean opposite). In other words, and following good-old-fashioned Uexkullian zoology: treat the animal as a subject. This requires that the means of studying an organism respect the possibility to understand it. Otherwise, neuroscience becomes neural science: the study of the properties of neural tissue and the side-effects on changes at the macroscopic level under the moniker of "behavior" — still fascinating yet uncoupled from its significance within the whole, which is the living organism[50]. Deliberate zoomorphism is preferable to unconscious anthropomorphism, which is often a kind of technomorphism — our tools do not tell us what things are or how they work; they enable us to answer properly-posed questions about that.

More generally, circular causality exemplifies the essential role of top-down causation[51] in biology (a huge and central topic which we cannot cover here). However, the behavior-to-circuits reduction reinforces the idea that causality's arrow can only be upwards, at the same time that it encourages the belief that the real thing is happening inside, not outside

---

[48] «I would say that a system is teleological if it has a mechanism which enables it to maintain a specific property despite environmental changes. (…) It must have certain types of compensating mechanisms — what we call essentially a negative feedback. (…) I would not say that a simple pendulum which moves in such a way that it strives to achieve the lowest potential energy is a teleological system. There are no compensating effects in the pendulum. Hence, as a system, it does not have an internal structure that enables it to compensate for environmental changes.» (**Nagel 1965**)

[49] «rather than concentrating on what an animal is doing, what may be more relevant is what the animal is trying to perceive. This double inversion (from the experimenter's point of view to the animal's and from action to perception) has three potentially critical implications for the study of cognition: first, motor output is a side effect of perceptual control and therefore quantitative data collection data is insufficient by itself, second, averaging across individuals may smear out control variables, and third, restrained experimental setups may not let animals control the relevant inputs; freedom in the requisite dimensions is essential. In short, control (circular causality) and subjectivity (animal centrism) may be essential ingredients in behavioral and cognitive neuroscience.» (**Gomez-Marin and Mainen 2016**)

[50] «Could not the application of physico-chemical methods possibly mean, in principle, such a destruction of the organism... Can it really teach us something about the functioning of the organism?» (**Goldstein 1934** p109)

[51] «five essentially different classes of top-down influence can be identified, and their existence demonstrated by many real-world examples. They are: algorithmic top-down causation; top-down causation via non-adaptive information control, top-down causation via adaptive selection, top-down causation via adaptive information control and intelligent top-down causation.» (**Ellis 2012**)

(as we mentioned at the beginning of our discussion, neuroscience is in danger of replacing 20th century behaviorism with 21st century technology-reinforced neuralism).

Circular causality implies that the behavior of animal is, to a great extent, self-caused. An important source of confusion may lie in conflating input-and-output with cause-and-effect. Note that when there is no distinction between cause and effect (circular causality), there can still be a difference between input and output (something comes in, something comes out — and in again, and out again). Contrary to "chain-like" thinking, effects are simultaneous with causes, not instantaneous.

One can turn this whole contradiction into complementary opposition: lineal causality is a particular case of circular causality upon breaking the loop[52]. But note that even if we artificially break the loop in the lab, the animal is still a creature in closed-loop. There is no escape from this: the challenge is to understand and apply the notion of causation in closed-loop (and then to realize its asymmetry: that the animal controls its perception of the environment more effectively than the world controls its behavior. It is remarkable how, during he 20th century, the scientific study of behavior and its neural underpinnings exploited the idea that animal behavior could and should be studied precisely by *not* letting animals behave...

Already 120 years ago, Dewey, in a stroke of genius, brought "the effect of output on input" to its ultimate consequences. He was able to articulate the notion of "circular act" by realizing that «the so-called response is not merely to the stimulus; it is *into* it» (**Dewey 1896**). In other words, he saw that «the expression of every impulse stimulates other experiences and these react into the original impulse and modify it» (**Dewey 1894** p14). Foreseeing the pioneering work of the cybernetics movement (which, paradoxically, missed its own point about "control in the animal" by concentrating on "information in the machine"), he asked «What shall we term that which is not sensation-followed-by-idea-followed-by-movement» (**Dewey 1896**), thus anticipating the idea of «back-reference of an experience to the impulse which induces it» (**Dewey 1894** p15). Dewey restored the conceptual equilibrium between movement and sensation[53]. In the same year (and in the same book!) we find a very similar pioneering account of perception and action as alternating, rather than alternative, views[54].

---

[52] «We have a general manipulative technique for making anything hot: we put it on a fire.» (**Wright 1971**)

[53] «The real beginning is with the act of seeing; it is looking and not a sensation of light. (...) In other words, we now have an enlarged and transformed coordination; the act is seeing no less than before, but it is now seeing-for-reaching purposes. There is still a sensori-motor circuit, one with more content or value, not a substitution of a motor response for a sensory stimulus. (...) it is a seeing-of-a-light-that-means-pain-when-contact-occurs.» (**Dewey 1896**)

[54] «In relation to each other inside the act of attention, most discussions of the subject appear to make the ear process merely a stimulus to which the hand adjustment is merely a response. But the question arises, What holds the ear to its work? Why does the reagent maintain his listening attitude? It may be replied that it is 'because he is told to." But he is not told to listen any more than he is told to move his hand. If the telling suffices in one case it should in the other. Moreover, he is not merely to listen, or even to listen just for the click, but to listen for the click as a pressure signal. It is this character of the click as a signal for pressure that keeps up the interest in it and the attention to it. (...) The hand therefore is stimuli as well as response to the ear, and the latter is response as well as stimulus to the hand. Each is both stimulus and response to the other. The distinction of stimulus and response is therefore not one of content, the stimulus being identified with the ear, the response with the hand, but one of function, and both offices belong equally to each organ. (...) it must be kept in mind that this latter is a distinction falling inside the act, not between the hand movement considered as the act, and the sound considered as its external

A notable exception to the usual conception of causation is Whitehead's philosophy of organism which, by abandoning «the notion of an actual entity as the unchanging subject of change» (**Whitehead 1929** p29), goes beyond the rational exercise of determination of successors by antecedents[55]. The implications of Whitehead's process ontology in the study of brain-behavior relationships would require a whole chapter by itself, if not a whole book. We cannot underscore enough here its great importance as a solid foundation to erect an organism-centric scientific study of topics that concern neuroscience in particular, and biology in general. Much more to be said in the future.

Causal-manipulative approaches are prone to reductionist and mechanistic frames. However, a re-elaboration of the notion of mechanism, with a tamed flavor of reductionism, has been proposed: it does without a decomposition of the system under study into "parts" and "interactions" and, instead, defines a mechanism in terms of "entities" together with their "activities"[56]. Such a neo-mechanistic approach is based on (and at the same time distinguished from) a blend of substance and process ontologies[57]. One may regard it as a smooth transition from the static notion of mechanism to something more akin to process. Yet, just because one can draw an arrow from A to B, it does follow that the arrow is a process-explanation (quite the contrary). Regrettably, in order to avoid the full implications of process philosophy, such a mingle becomes dualistic at best and self-contradictory at worst[58].

Let us briefly mention quantum causality, where causes can be non-local to their effects. «The inference to (efficient) causation is always based on an *interpretation* of observed correlations (...), correlations for which both types of classical efficient causation can be rule out, even experimentally» (**Atmanspacher 2014**). Entanglement (both truly quantum and also classical, by means of epistemic inaccessibility) remains an alternative as intriguing as interesting: «this does not mean that the correlations have no reason at all or are just 'causeless' (...). But their cause is not of the efficient variety – in Aristotle's terminology, it comes closest to the notion of formal causation.» (**ibid**)

To end, let us insist on the following essential truism: the study of the behavior of organisms is crucial in order to understand the behavior of organisms[59].

---

stimulus or "cause." In a word, the reagent reacts as much with his ear as he does with his hand.» (**Angell and Moore 1896** p252)

55 «The concrescence of each individual actual entity is internally determined and is externally free». More explicitly: «The doctrine of the philosophy of organism is that, however far the sphere of efficient causation be pushed in the determination of components of a concrescence (...) beyond the determination of these components there always remains the final reaction of the self-creative unity of the universe. This final reaction completes the self-creative act by putting the decisive stamp of creative emphasis upon the determinations of efficient cause.» (**Whitehead 1929** p47)

56 «Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions» (**Machamer, Darden and Craver 2000**)

57 «We meant thereby to distinguish our position, or way of thinking, from a substance ontology and from process ontology, and we chose "entity" and "activity" because these terms seemed to carry fewer historical and philosophical presuppositions than "substance" and "process".» (**Machamer 2004**)

58 «Process philosophers would have us redefine all entities in terms of combining processes, but this seems a bit too strange. Therefore, we (MDC 2000) decided to be dualist.» (**ibid**)

59 And this shall include, together with theory, the appreciation for descriptive science: «Simply describing what we see is not considered very scientific nowadays and 'descriptive science' has become a derogatory

Part 4. FINALE: know what you mean and say it

I took the liberty and challenge of not speaking to the choir nor spitting to the noir[60]. This critique is certainly a "minority report", and it may be deemed too abstract and philosophical for the mainstream neuroscientist. But science is not only about data and tools, but primarily about the concepts that such data and tools allow to probe[61].

Here it was deemed necessary to halt and ask, in the midst of such great efforts in manipulating neural circuits, to what extent they *explain* of behavior. Deconstructing the most common phraseology found in neuroscience, we identified, traced and articulated the primacy of some basic assumptions about what explanation is and what behavior is. We refrained from opposing neural chauvinism with behavioral chauvinism.

We followed the original call for understanding and realized that the strength in the initial notion of explanation runs down a gradient of concessions and approximations until we are left with hardly more than *counterfactual dependence*. Overall, we concluded that neural circuit necessity and sufficiency may not be necessary nor sufficient to explain behavior. Paraphrasing Woese, there is nothing wrong with N&S per se. Wrong is when N&S comes to define explanation neuroscience.

Neuroscience is a vibrant endeavor, yet it is hard to see where we are going with what we are doing, what we mean by understanding, or what counts as a true insight. From cellular neurobiology to cognitive neuroscience, no matter how disparate their interests and different their communities, we all ultimately subscribe, one way or another, to the "neurons explain behavior" mantra. Trying to bring clarity and honesty about what we *can't* currently explain —and most likely *won't* be able to solve— with our methodological procedures and conceptual frameworks is one more reminder for humility, after so much pride. One thing is clear: the amount of understanding we can gain about the neural *implementation* of behavioral processes depends on the quality of prior investigations of such behavioral processes, and also on the conceptual grounding that bridges the relation between neural and behavioral levels of organization.

---

term. We must have a hypothesis to test, or better, a controlled experiment that can be performed to identify ecological rules and laws. However, if "ecological rules" were followed by all systems, unexpected things would not happen, which is evidently not the case. Deviations from rules are the main determinants of history, but we cannot test something that is unexpected. As such, our quest to identify rules and regularities could be preventing us from understanding the history of these systems. Paradoxically, we aim at understanding historical systems while using ahistorical approaches! We need a means of reporting these contingencies so that we can better understand the historical trajectory of ecological systems.» (**Boero 2013**)

[60] «Both the criteria of plausibility and of scientific value tend to enforce conformity, while the value attached to originality encourages dissent. This internal tension is essential in guiding and motivating scientific work. The professional standards of science must impose a framework of discipline and at the same time encourage rebellion against it. They must demand that, in order to be taken seriously, a investigation should largely conform to the currently predominant beliefs about the nature of things, while allowing that in order to be original it may to some extent go against these.» (**Polanyi 1962**)

[61] «Science is impelled by two main factors, technological advance and a guiding vision. A properly balanced relationship between the two is key to the successful development of a science: without the proper technological advances the road ahead is blocked. Without a guiding vision there is no road ahead; the science becomes an engineering discipline, concerned with temporal practical problems.» (**Woese 2004**)

It is surprising how little effort is made (by scientists) to explain what is actually meant by (scientific) explanation. Interwoven attitudes reflect ingrained habits of the intellect: the idea that pristine observation, when combined with powerful manipulation, shall disclose the truth about things — even when the pursuit of truth is actually deemed as impossible in principle and irrelevant in practice. Following Bacon's dictum (and in contrast with Aristotle's) one should "torture" *nature* until she spits out her answers. Paradoxically, the force required in such an interrogatory (engineering new tools, analyzing big data, trading necessity-sufficiency tests for explanation, selling our findings, getting grants, etc) drains the capacity to consciously recall what we wanted to ask *her* in the first place. This sort of scientific amnesia can only perpetuate itself unless we are willing to equip ourselves with the kind of literacy that shall make precise the difference between what we do versus what we claim we do. In sum, *know what you mean and say it.*

Acknowledgements. I thank Björn Brembs, André Brown, Adam Calhoun, Asif Ghazanfar, José Gomes Pinto, Gordon Globus, Eyal Gruntman, Rod Hemsell, Johannes Jaeger, Konrad Kording, Gonçalo Lopes, Adam Matic, Laura Navío, Joana Rigato, Troy Shirangi, and Ibrahim Tastekin, for feedback. My views need not coincide with theirs.

References

Angell, J. R., & **Moore**, A. W. (**1896**). Studies from the psychological laboratory of the University of Chicago: I. Reaction-time: A study in attention and habit. *Psychological Review*, *3*(3), 245.

Anscombe, G. E. M. (**1971**). Causality and Determination. In E. Sosa M. Tooley (ed.), Causation. 1993 Oxford p88-104.

Ashby, W. R. (**1958**). Requisite variety and its implications for the control of complex systems. *Cybernetica* 1:2, p. 83-99

Atmanspacher, H. (**2014**). Roles of causation and meaning for interpreting correlations. Journal of Analytical Psychology, 59(3), 429-434.

Beer, C. G. (**1980**). Perspectives on Animal Behavior Comparisons. In *In Comparative Methods in Psychology* (pp. 17-64). Erlbaum Hillsdale, New Jersey.

Bell, H. C. (**2014**). Behavioral variability in the service of constancy. *International Journal of Comparative Psychology*, *27*(2).

Von **Bertalanffy**, L. (**1950**). The theory of open systems in physics and biology. *Science*, 111(2872), 23-29.

Boero, F. (**2013**). Observational articles: a tool to reconstruct ecological history based on chronicling unusual events. F1000Research, 2:168.

Bunge, M. (**2009**). Causality and modern science. Fourth Revised Edition, Transaction Publishers, New Jersey.

Cools, A. R. (**1985**). Brain and behavior: hierarchy of feedback systems and control of input. In *Perspectives in ethology* (pp. 109-168). Springer US.

Cziko, G. A. (**2000**). *The Things We Do: Using the Lessons of Bernard and Darwin to Understand the What, How, and Why of Our Behavior*. The MIT Press.

Dawkins, R. (**1976**). Hierarchical organisation: a candidate principle for ethology. In: Growing Points in Ethology (Ed. by P. P. G. Bateson & R. A. Hinde). Cambridge University Press.

Dewey, J. (**1896**). The reflex arc concept in psychology. *Psychological review*, *3*(4), 357.

Dewey, J. (**1894**). *The study of ethics: A syllabus*. Register Publishing Company.

Dewey, J. (**1930**), Conduct and Experience. In: Psychologies of 1930 (Ed. By C. Murchison). The International University Series in Psychology. Worcester, Massachusetts. Clark University Press.

DiDomenico, R., & **Eaton**, R. C. (**1988**). Seven principles for command and the neural causation of behavior. *Brain, behavior and evolution*, *31*(3), 125-140.

Ellis, G. F. (**2011**). Top-down causation and emergence: some comments on mechanisms. *Interface Focus*, rsfs.2011.0062.

Eaton, R. C., & **DiDomenico**, R. (**1985**). Command and the Neural Causation of Behavior: A Theoretical Analysis of the Necessity and Sufficiency Paradigm; pp. 149–164. *Brain, behavior and evolution*, *27*(2-4), 149-164.

Forscher, B. K. (**1963**). Chaos in the brickyard. Science, 142(3590), 339.

Goldstein, K. (**1934**). The Organism, New York: Zone Books

Gomez-Marin, A., & **Mainen**, Z. F. (**2016**). Expanding perspectives on cognition in humans, animals, and machines. *Current opinion in neurobiology*, *37*, 85-91.

Hogan, R. (**2001**). Wittgenstein was right. *Psychological Inquiry*, Vol 12(1) 27.

James, W. (**1890**). The principles of psychology. 1950 New York: Dover.

Jankélévitch, V. (**2015**). *Henri Bergson*. Duke University Press.

Jonas, E., & **Kording**, K. (**2016**). Could a neuroscientist understand a microprocessor? bioRxiv 10.1101/055624.

Karmon, A., & **Pilpel**, Y. (**2016**). Biological causal links on physiological and evolutionary time scales. *Elife*, *5*, e14424.

Kortmulder, K. (**1998**). Play and evolution: Second thoughts on the behavior of animals. International Books, The Netherlands.

Lazebnik, Y. (**2002**). Can a biologist fix a radio?—Or, what I learned while studying apoptosis. *Cancer cell*, *2*(3), 179-182.

Machamer, P. (**2004**). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science*, *18*(1), 27-39.

Machamer, P., **Darden**, L., & **Craver**, C. F. (**2000**). Thinking about mechanisms. *Philosophy of science*, 67 (1):1-25.

Marken, R. S. (**2001**). Controlled variables: Psychology as the center fielder views it. *The American journal of psychology*, *114*(2), 259.

Maturana, H. R. (**1995**). Biology of self-consciousness. In *Consciousness: Distinction and reflection* (pp. 145-175). Bibliopolis.

Medawar, P. B. (**1978**). Is the scientific paper fraudulent? Yes; it misrepresents scientific thought. Science in Books SR/August 1:42-43.

Merleau-Ponty, M. (**1942**). The Structure of Behavior. trans. by A. Fischer. Boston 1963, Beacon Press.

Nagel, E. (**1965**). Types of causal explanation in science. In *Cause and Effect*, ed. Daniel Lerner. The Free Press, New York.

Noë, A. (**2004**). *Action in perception*. MIT press.

Overton, W. F. (**2006**). Developmental psychology: Philosophy, concepts, methodology. In *Handbook of child psychology*, Vol 1. ed. Richard M. Lerner. John Wiley & Sons, Inc. New Jersey.

Polanyi, M. (**1962**). The republic of science: Its political and economic theory. *Minerva* 1:54-74.

Powers, W. T. (**1973**). *Behavior: The control of perception*. Chicago: Aldine.

Powers, W. T. (**1998**). About Stimulus Response Theory and Perceptual Control Theory". Post to the Control Systems Group Network.

Rosenblueth, A., **Wiener**, N., & **Bigelow**, J. (**1943**). Behavior, purpose and teleology. *Philosophy of science*, 10(1), 18-24.

Russell, B. (**1913**). On the notion of cause. *Proceedings of the Aristotelian Society*, New Series, Vol. 13, pp. 1-26.

The Edwin Smith Surgical Papyrus (**1930**). By James H. Breasted. Chicago: University of Chicago Press.

Tinbergen, N. (**1963**). On aims and methods of ethology. Zeitschrift für Tierpsychologie, 20(4), 410-433.

von **Wright**, G. H. (**1971**). Explanation and Understanding. Cornell University Press, *New York*.

De **Waelhens**, A. (**1942**). A Philosophy of the Ambiguous, Foreword to the Second French Edition of The Structure of Behavior by Merleau-Ponty. Boston: Beacon Press

Whitehead, A. N. (**1929**). Process and reality. 1978. *Corrected ed. Ed. David Ray Griffin and Donald Sherburne. New York.*

Whitehead, A. N. (**1933**). Adventures of Ideas. 1967. The Free Press. New York.

Woese, C. R. (**2004**). A new biology for a new century. *Microbiology and molecular biology reviews*, *68*(2), 173-186.