AVANT, Vol. VI, No. 2/2016
ISSN: 2082-6710 avant.edu.pl/en
DOI 10.26913/70202016.0112.0003

# Integrating cognitive (neuro)science using mechanisms

**Marcin Miłkowski**
Institute of Philosophy and Sociology,
Polish Academy of Sciences
*mmilkows@ifispan.waw.pl*

## Abstract

Abstract: In this paper, an account of theoretical integration in cognitive (neuro)science from the mechanistic perspective is defended. It is argued that mechanistic patterns of integration can be better understood in terms of constraints on representations of mechanisms, not just on the space of possible mechanisms, as previous accounts of integration had it. This way, integration can be analyzed in more detail with the help of constraint-satisfaction account of coherence between scientific represen-tations. In particular, the account has resources to talk of idealizations and research heuristics employed by researchers to combine separate results and theoretical frameworks. The account is subsequently applied to an example of successful integration in the research on hippocampus and memory, and to a failure of integration in the research on mirror neurons as purportedly explanatory of sexual orientation.

**Keywords**: theoretical integration; mechanistic explanation; LTP; interfield theories; constraint-satisfaction.

In this paper, I defend an account of integration in cognitive (neuro)science from the mechanistic perspective. I argue that previously proposed mechanistic patterns of integration can be better understood in terms of constraints on representations of mechanisms, not just on the space of possible mechanisms. This way, integration can be analyzed in more detail with the help of constraint-satisfaction account of coherence between scientific representations. In particular, the account has resources to talk of idealizations and research heuristics employed by researchers to combine separate results and theoretical frameworks.

To illustrate this proposal, two examples will be outlined: (1) research on the hippocampus and memory consolidation, which is a classical illustration of multi-level research, according to Craver (2002; 2005); (2) a recent proposal to understand sexual orientation as related to the mirror-neuron system (Ponseti et al. 2006), which is an instance of a clear failure at inter-level integration (and more).

## 1. Mechanistic integration, Craver and Darden style

In this section, I will introduce the neomechanistic account of explanation and show how it relates to issues of integration and unification, particularly, in cognitive (neuro)science. What is notable is that the neomechanistic account does not stress the theoretical autonomy of psychology or cognitive science vis-à-vis neuroscience; in contrast, it argues for integrating research that spans multiple fields, without calling for a ruthless elimination of any of them. Then I will briefly describe the extant taxonomy of mechanistic integration as proposed by Craver and Darden.

According to the neomechanistic account of explanation, to explain a phenomenon $\phi$ is to elucidate the causal structure of the mechanism that gives rise to $\phi$ (Machamer, Darden, and Craver 2000; Glennan 2002; Bechtel and Abrahamsen 2005). While there are multiple definitions for the term *mechanism*, the core idea is that a mechanism is an organized system that comprises causally relevant component and operations (or activities). Component parts of the mechanism interact, and their organized operation contributes to the capacity of the mechanism to exhibit $\phi$. (For a recent review, see, e.g., [Illari and Williamson, 2011]).

The mechanistic account of explanation is privy to issues of inter-field research (Darden and Maull 1977; Craver and Darden 2013). Inter-field theories relate to at least two fields of study. By a “field of study”, Darden and Maull mean “an area of science consisting of the following elements: a central problem, a domain consisting of items taken to be facts related to that problem, general explanatory factors and goals providing expectations as to how the problem is to be solved, techniques and methods, and, sometimes, but not always, concepts, laws and theories which are related to the problem and which attempt to realize the explanatory goals”[13] (Darden and Maull 1977, 44). Two fields may appeal to the same spatiotemporal locations, entities, or activities, and one may provide a better understanding of the spatiotemporal relationships, causal relationships, physical nature, structure, or function thereof.

[13] As Bechtel (1986, 11–13) notices, the central problems may be solved over time, which does not mean that the field or discipline is going to disappear; the fields should be therefore defined by a certain genealogy of problems rather than a single central problem.

However, in looking at cognition, it is clear that cognitive processes may be explained in different ways by various disciplines.

The practice, methodology and theory of cognitive science do not rest on autonomy from neuroscience. It is the opposite, as the boundary between cognitive science and cognitive neuroscience is blurred. Defenders of the new mechanistic approach argue the functionalist explanations of cognitive capacities are essentially incomplete explanations of mechanisms (Piccinini and Craver 2011). These are incomplete because they do not cite all causally relevant factors for their explananda. Others have pointed out that functional analyses may lead to positing entities or activities devoid of any causal relevance for the phenomenon (Miłkowski 2013, chap. 3), because the functional analysis yields factors sufficient for the functional capacity to appear (Cummins 1975). These factors may not be at play in a mechanism; hence, they may be causally irrelevant. Functional analyses may (1) fail to describe actual causal factors; (2) posit irrelevant or epiphenomenal factors instead. In contrast, mechanistic explanations reject the functionalist assumption that factors that are logically sufficient for the explanandum to occur are genuinely explanatory. They need to be further constrained and become actual causal factors. This requires, arguably, meshing cognitive science with neuroscientific research. Some even claim there has been a silent revolution that turned cognitive science into cognitive neuroscience (Boone and Piccinini 2015).

For example, well-established results in neuroscience constrain psychological model-building; this is how one can define the role of double dissociations (Glymour 1994). While deemed highly controversial (Van Orden, Pennington, and Stone 2001), double dissociations help to delineate individual mechanisms. Theorizing about different aphasias and underlying brain lesions is arguably mutually constraining, even if one cannot identify whole brain mechanisms to be juxtaposed.

The new mechanistic approach does not render cognitive science obsolete vis-à-vis neuroscientific or molecular explanations as other “ruthlessly reductive” accounts would have it (Bickle 2003). Instead, some argue that mechanisms are not made, in any sense, unreal when underlying factors are introduced (Craver 2005). While it may be still argued the new mechanistic approach is, in an important sense, reductive, as it appeals to some reductive heuristics in the search for explanations (Hensel 2013), these heuristics are not supposed to supplant more abstract explanations used in genuinely mechanistic explanations (Levy and Bechtel 2013). While mechanists stress the importance of interfield integration, they do not argue for *replacement* of some fields by other fields that study phenomena at smaller spatiotemporal scales.

Cognitive science comprises multiple fields that contain stronger and weaker connections among them. The stronger the connection is between fields A and B, the higher the probability the models will integrate insights from both A and B. In this paper, by a *model of a mechanism* M, I mean a scientific representation of M, be it verbal, diagram-like, computational, or purely formal. Models are cognitive artifacts employed by researchers to describe mechanisms, and this usage corresponds roughly to how the term *model* is used in cognitive science. *Integration*, in this context, refers to the combination of different models. But integration may not lead to complete *unification*. Unification results from developing general, simple, elegant, and beautiful explanations. Most published models in cognitive science are minimal—they focus on individual tasks, rather than all features of cognitive systems. Newell (1973) observed this may lead to disintegration and fragmentation as the stress on minimal models may create disunity. To avoid this, Allen Newell proposed redirecting the focus of research to cognitive architectures (Newell 1990).

Cognitive architectures are hypotheses about the general organization of cognitive mechanisms, and according to Newell, they seek to achieve theoretical unification. However, unification may occur on a much smaller scale. Unified explanations are intuitively assessed as simple, general, and beautiful, but they need not be the simplest, most general, and most beautiful. The appeal to aesthetic criteria may seem to invoke non-analyzable, elusive properties, and maybe this is the reason unification has rarely been analyzed by the defenders of mechanistic explanation. Providing detailed elucidations of these properties goes beyond the scope of the current paper, as my focus is on integration and building larger integrative models that span multiple different fields (but see Miłkowski 2017). Unification in cognitive science may proceed (i) via schemes of structures or (ii) via elementary processes (Danks 2014, 176). Scheme-centered unification "arise when we have a collection of distinct cognitive theories and models that are nonetheless all instantiations of the same type of structure (in some sense)." In other words, schema-centered accounts argue for cognitive unification "in virtue of some common template that is shared by all the individual cognitive models, rather than through shared cognitive elements (representations, processes, or both) across those models." Process-based unifications try "to show how coherent cognition arises from shared processes, where those processes are typically small building blocks that combine to yield complex cognition" (ibid.).

Let me put unification to side then and turn to integration. Darden and Craver (2013, chap. 10) have identified at least three ways mechanisms—and fields[14]—may become integrated. Clearly, listing them is not supposed to provide neither a systematic taxonomy of all the possible ways mechanisms can be integrated, nor a detailed understanding of what integration is. However,

[14] Fields may be integrated in non-mechanistic ways (for example, when they do not rely on mechanistic explanations), and their integration may depend on further sociotechnical factors. I do not deal with the integration of fields as such in this paper. For a broader treatment of integration, including interfield integration, see (Bechtel 1986).

these three patterns are all based on simple juxtapositions or part-whole relationships, and this is probably what makes them particularly easy to observe and understand.

*Simple integration* puts models of mechanisms together as pieces of a puzzle that fit together. Two fields may simply study cognition in a similar way, but may emphasize different mechanisms. For example, separate stages of visual processing may be studied in relative isolation and then combined to produce a larger mechanism. Notice that, in this case, all integrated models are on the same level of mechanistic organization, which shows simple integration is not inter-level. The notion of the level is understood in the mechanistic framework in terms of proper part-whole relationships: whatever is a proper part of a larger mechanism is at the lower level of its organization (Craver 2007). Hence, simple integration occurs when no mechanism being integrated is a proper part of another.

*Inter-level integration* occurs when another level of organization is added to complete an explanation: a larger mechanism may be proposed that contains the previous one, or a lower-level mechanism is added as a proper part of the previous mechanism. This integration usually encompasses the development of an existing explanation of a phenomenon by providing an underlying mechanism of the phenomenon (Thagard 2007). For example, my ability to recognize a certain visual pattern may be identified as a capacity of a mechanism in the visual cortex.

*Inter-temporal integration* puts one mechanism in the context of another mechanism that functions on another time-scale. Here, one mechanism becomes a temporal proper part of another mechanism. Think of developmental mechanisms shaping my perceptual capacities, for example, the ability to see colors.

These three patterns of integration can be easily observed in cognitive science. They correspond to spatial and temporal adjacency (simple integration) and spatial or temporal containment (inter-level and inter-temporal integration). What is supposed to be at play in research practices is more than these simple relationships. Craver has argued “different fields integrate their research by adding constraints on a multilevel description of a mechanism” (Craver 2005, 373). In the next section, I will develop his account and propose a slight modification in the notion of constraint used to analyze inter-field integration.

## 2. Mechanistic integration as constraint satisfaction

In this section, I elucidate the notion of *constraint* in the mechanistic literature, which has been framed in ontic terms. I argue for a slight modification: constraints that lead to integration (and, possibly, other kinds of methodological constraints) can be accounted for, more generally, if one considers their role in shaping scientific representations. In doing this, I refer to previous work on representational constraints and show how Thagard's account of constraint satisfaction can shed light on interfield constraints.

Craver defines his notion of *constraint* in the following way:

> A constraint is a finding that either shapes the boundaries of the space of plausible mechanisms or changes the probability distribution over that space (that is, the probability that some point or region of the space accurately describes the actual mechanism). Some constraints exclude regions of the space; they show that some set of possible mechanisms is impossible given what is known about the components and their organization … Constraints on the space of possible mechanisms, in short, constitute the relevant evidence for evaluating how-possibly descriptions of mechanisms. Progress from how-possibly to how-actually descriptions of a mechanism can thus be conceived as a process of shaping and constricting the space of plausible mechanisms. (Craver 2007, 247–48).

However, some specific epistemic constraints do not seem appropriately characterized as *findings*. Craver (2007, 26) lists several methodological principles used by Wesley Salmon as *constraints* on the space of possible explanations:

> (E1) mere temporal sequences are not explanatory (temporal sequences);
>
> (E2) causes explain effects and not vice versa (asymmetry).

It seems more appropriate, therefore, to talk of some scientific representations as (mutually) constraining, as (E1) or (E2) is more appropriately described as a methodological norm or principle rather than a finding. A causal model of a mechanism should be constrained by (E1) to be fully explanatory. One could also conceive several models of mechanisms as constraining some resultant overall model, which serves as their mechanistic amalgamation. These models may contain not only empirical findings, but also some methodological principles or methods (e.g., representational conventions used to create computer simulations).

Note, for example, one of the constraints mentioned by Craver (2007) is mutual manipulability. This constraint appeals to causal considerations, which cannot be easily rephrased in terms of constraints understood as *findings*: it should be the case that entities and activities are appropriately mutually manipulable (that is, support bottom-up and top-down interventions), but the constraint is *not* a finding or discovery in any usual sense.

One obvious way to evaluate my proposal to frame constraints as operating on representations is to compare ontic and representational accounts of constraints. Recently, David Danks has proposed a representational account of constraints to analyze inter-theoretic and inter-model relationships. This is not mere terminological similarity but, arguably, an attempt to elucidate the same concept: “one theory *S* constrains another theory *T* if the extent to which *S* has some theoretical virtue *V* (e.g., truth, predictive accuracy, explanatory power) matters for the extent to which *T* has *V*” (Danks 2014, 31). This means that, if *S* constrains *T* because of a certain theoretical virtue, then if we care about this virtue in *T*, then we should care about it in *S*. Different virtues give rise to different constraints.

In Danks’s analysis, constraints act upon representations, while Craver assumes these operate on the space of possible mechanisms. This corresponds to the distinction between epistemic and ontic accounts of explanation. While the ontic and epistemic talk are largely interchangeable, there are notable differences. There may be some theoretical virtues related to idealization, and idealization introduces non-veridical scientific representations, which may be distorted descriptions of mechanisms. Constraining the space of possible mechanisms to find the idealized mechanism may be a futile endeavor, just because some idealizations may introduce entities or activities that are, strictly speaking, physically impossible, such as mass points or frictionless motions. The space of possible physical mechanisms would then be empty. One solution is to look at the space of logically possible mechanisms, but it is not at all clear whether merely logically possible mechanisms explain anything *ontically* (and some idealizations are, strictly speaking, even logically impossible). Instead, one could talk of constraints over the space of descriptions of mechanisms and hold that some parts of descriptions are not supposed to be explanatory, per se, but are indispensable in explanations for various reasons. Building idealized scientific representations may be justified, regarding their theoretical virtues, but is not best accounted for in ontic terms.

A similar argument may be formulated, regarding other representational virtues. One of the simplest axiomatizations of the propositional calculus offered by Jan Łukasiewicz in his notation is *EEpqEErqEpr*. The axiom is obscure even to those trained in Reverse Polish Notation (*E* stands here for binegation, should you wonder). Similarly, the most parsimonious representations may be cumbersome to use in scientific practice. So, one may require that the model of the mechanism be easy to use in a given scientific practice or that integrated models should be perspicuous. These requirements are easy to frame as theoretical principles constraining the representation.

So, to sum up, the difference between the representational account of constraints, which explicitly appeals to theoretical virtues and the ontic account offered by Craver is that the first is simply easier to apply to scientific representations in case of idealizations or other representational ideals, but can easily express the ontic requirements. This account is sufficiently similar to fit the mechanistic framework of explanation.

The weakest constraint described by Danks is a *truth-constraint*: two bodies of knowledge satisfy a truth-constraint if they can be both true at the same time. The received account of inter-theoretical reduction involves truth-constraints, as Danks argues. However, truth-constraining is a weak relation of logical coherence. The wave theory of light does not exclude the particle theory of light, so they satisfy the truth-constraint, even though they involve a different account of the basic nature of light. Basically, any mechanistic model M1 that does not deny another, M2, may be endorsed at the same time, and as far as M1 and M2 can be true, they would be truth-constrained, but not mechanistically integrated.

Before I discuss mechanistic constraints deeper, it's important to note another advantage of framing constraints in representational terms. Interestingly, applying the truth-constraint is a case of attaining logical coherence between representations, which is one of the kinds of coherence analyzed by Paul Thagard (2000). Thagard also uses the term *constraint*, and even if he offers no normal definition of this term, his contextual definition of constraint in terms of coherence is compatible with Danks's account. Two kinds of constraints are introduced with two reductive definitions: "If two elements cohere, there is a positive constraint between them. If two elements incohere, there is a negative constraint between them" (Thagard 2000, 17). Thus, the ontic constraints that Craver describes as excluding some regions of the space of possible mechanisms are simply negative constraints. The positive constraints drive the search for the actual mechanism responsible for some explanandum phenomenon in the space of possible mechanisms.

Constraints may also be stronger or weaker, and attaining coherence between representations requires solving the *coherence problem*, defined this way:

> Let *E* be a finite set of elements {$e_i$} and *C* be a set of constraints on *E* understood as a set {($e_i$, $e_j$)} of pairs of elements of *E*. *C* divides into *C*+, the positive constraints on *E*, and *C*-, the negative constraints on E. With each constraint is associated a number *w*, which is the weight (strength) of the constraint. The problem is to partition *E* into two sets, *A* and *R*, in a way that maximizes compliance with the following two coherence conditions:
> • If ($e_i$, $e_j$) is in *C*+, then $e_i$ is in *A* if and only if $e_j$ is in *A*.
> • If ($e_i$, $e_j$) is in *C*-, then $e_i$ is in *A* if and only if $e_j$ is in *R*.
> Let *W* be the weight of the partition, that is, the sum of the weights of the satisfied constraints. The coherence problem is then to partition *E* into *A* and R in a way that maximizes *W*. Because *a coheres with b* is a symmetric relation, the order of the elements in the constraints does not matter (Thagard 2000, 18).

Thagard (2000, 28) analyzes the complexity of algorithms that might solve coherence problems, which are generally NP-complete, or non-tractable for practical purposes for non-trivial domains. There are, however, some good approximations based on heuristics, including a connectionist constraint-satisfaction algorithm preferred by Thagard. Even if most models of mechanisms in life sciences remain semi-formal, with diagrams and verbal descriptions being parts of complex scientific models, there are domains in which computational modeling is prominent. This includes cognitive science, in which around 80-90% of theoretical papers appeal to computational models (Busemeyer and Diederich 2010). This account of integration in terms of constraint satisfaction can also offer a practicable solution.

If positive constraints and negative constraints may operate on models of mechanisms and on methodological principles, then an appropriately defined negative constraint may exclude models that do not satisfy methodological principles of causal explanation. Just because there are algorithms for finding causal relationships in the experimental data (Glymour 2001; Spirtes, Glymour, and Scheines 2000), it seems possible, at least in principle, to use them to constrain the models of mechanisms by filtering out the models that do not conform to observational data (see e.g., Triantafillou and Tsamardinos 2015).

Models in cognitive science, even if stated in a machine-readable form, are admittedly still rarely integrated automatically, especially if they are supposed to conform to semantic constraints that refer to spatial and temporal properties of entities and activities in mechanisms; integrating them remains more art than science. But the constraint-satisfaction account of integration for mechanistic models does not serve merely a practical purpose.

Constraints are not limited to truth values of models or their components (I have mentioned the causal constrain in passing). Otherwise, providing an integrated model of the mechanism would boil down to finding a minimal coherent model that satisfies this constraint. This is not the case, even for reduction in the classical sense (Nagel 1961). Danks (2014, 31) claims that "reduction is arguably the strongest possible inter-level constraint, so it can sometimes make sense to focus on finding reductions of some theory H." Things are not so simple. Reduction need not lead to a deep unification if the reducing theory $T_1$ is nothing but a language able to express another theory $T_2$ positing no substantial connections between its claims and the claims of $T_1$ (cf. Bechtel 1986, 41). Truth-constraints guarantee no kind of unification, and even in the strongest case, inter-theoretical reduction, coherent models may satisfy these constraints with no substantial inter-model connections.

In this context, it is important distinguish between mechanistic integration and non-mechanistic accounts of inter-model integration. As it relates to cognitive modeling, weak truth-constraints may only be applicable to the model's output or products of cognitive processes. Such modeling has been determined to offer a non-causal and non-mechanistic explanation (Irvine 2014). However, because most models of cognition aim at explaining cognitive processes, the integration offered by non-mechanistic explanations that focus merely on cognitive products, not processes, lacks depth and does not allow researchers to find common entities and processes in various models. Product-oriented integrations and unifications are not entirely satisfactory, and from the mechanistic point of view, they are incomplete, as they do not satisfy the completeness norm for mechanistic explanations. The completeness norm requires the model of the mechanism contain all causally relevant variables (Craver 2007). (However, it does not require specifying all possible details, and it also does not exclude idealization (Miłkowski 2016).)

Mechanistic constraints concern the entities and activities presupposed in integrated models, and this way, go beyond mere truth constraints. They are more of a semantic nature. Take the constraint that two integrated theories should appeal to the same entities: two theories of light no longer satisfy it, unless a unifying theory is proposed, for example, stating that light has both the nature of a particle and a wave at the same time. Theories of light are not mechanistic, as they do not explain phenomena with spatiotemporally delineated organized systems of components and operations. The constraints may be also more specific; for example, there are spatial constraints that concern the size, shape, location, connection, and compartmentalization of component entities; only some may be initially satisfied. Table 1 lists constraints on the space of possible mechanisms enumerated by Craver.

Table 1. *Intralevel and interlevel constraints on multilevel mechanisms* (Craver 2007)

- Intralevel constraints
    - Componency constraints
    - Spatial constraints
        - Size
        - Shape
        - Location
        - Connection
        - Compartmentalization
    - Temporal constraints
        - Order
        - Rate
        - Duration
    - Active constraints
- Interlevel constraints
    - Accommodative constraints
        - Top-down accommodation
        - Bottom-up accommodation
    - Spatial and temporal constraints
    - Mutual manipulability constraints

Three patterns of integration introduced in the first section can be naturally accounted for in terms of mechanistic constraints. With simple integration, a larger mechanism is simply posited, and it is constrained in that its model should contain entities and operations of two (or more) mechanisms that constitute it. The larger mechanism integrated in the inter-level way is constrained also to contain both mechanisms and, particularly, to have components of the first mechanism identified with the second mechanism. In the inter-temporally integrated mechanisms, it is primarily the operations of both mechanisms that are supposed to be identified with one another (or some causal relationship is supposed to obtain between these two mechanisms).

Craver stresses that integration is supported by multiple kinds of constraints. More complex integration can, therefore, go beyond the simple patterns we have seen before. But constraints in his sense are merely ontic, while, as we will see in the following sections, other kinds of constraints are used in integration efforts in neuroscience. The present account directly corresponds to considerations cited in the previous mechanistic accounts of integration and underlines the representational role of these constraints, that is on how they work on the descriptions or representations of mechanisms.

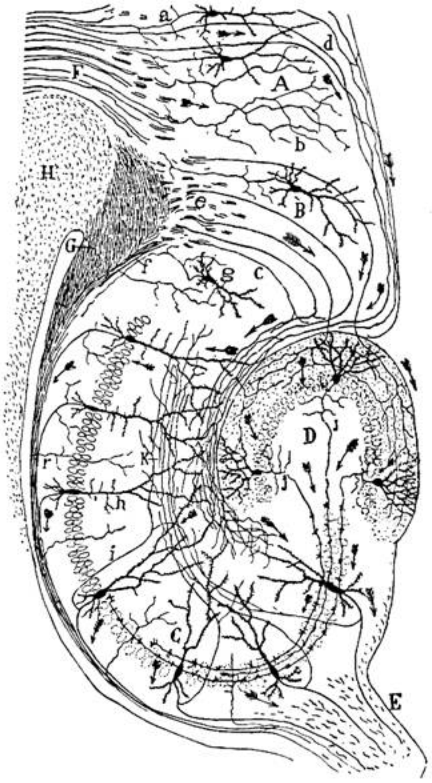
Just because the constraint satisfaction account can describe all previously found kinds of integration and more, it is a slightly more general proposal for a unifying account of mechanistic integration. But it remains to be shown that it is both descriptively and normatively adequate. It should be possible to describe integrated models of mechanisms via constraints, and good integrated models should be distinguished from bad. To demonstrate the account has these features, two cases studies will be examined.

The first is a relatively well-studied case of Long Term Potentiation (LTP) as involved in mechanisms of memory. It is an example of a successful integration of psychological constraints in neuroscientific theories. The constraint account is more plausible than simple inter-level integration. It will be also shown that the research program is constrained by some theoretical and methodological considerations, not just by findings related to mechanisms. The second case is an example of a failed integration—namely the case of the hypothesis of mirror neuron system (MNS) in the account of sexual orientation (or rather: in sexual arousal elicited by stimuli compatible with one's sexual orientation). I will argue it fails for multiple reasons, not only because the models of mechanisms posited for sexual arousal are not constrained mechanistically by the MNS, but also because methodological principles of causal explanation are not adhered to.

## 3. Hippocampus and memory

LTP is now believed to be a process of synaptic plasticity involved in mechanisms of memory, and the discovery of the mechanistic link between LTP and the hippocampus (Bliss, Gardner-Medwin, and Lømo 1973; Bliss and Gardner-Medwin 1973) was studied by Craver (2003; 2005; 2007), who opposed Bickle's (2003) ruthlessly reductive interpretation of this discovery. In the 1950s, the plasticity of the hippocampus (see Figure 1), which is a major component of the brains of human beings and other vertebrates, was experimentally confirmed (Green and Adey 1956). But plasticity was not yet linked to memory functions.

Figure 1: Hippocampus as drawn by Santiago Ramón y Cajal



The unsuccessful search for the dedicated mechanisms of memory, or the engram, was summarized influentially by Karl Lashley:

> This series of experiments has yielded a good bit of information about what and where the memory trace is not. It has discovered nothing directly of the real nature of the engram. I sometimes feel, in reviewing the evidence on the localization of the memory trace, that the necessary conclusion is that learn-

> ing just is not possible. It is difficult to conceive of a mechanism that can satisfy the conditions set for it. Nevertheless, in spite of such evidence against it, learning sometimes does occur (Lashley 1950).

Lashley's pessimism led researchers to assign other functions to the hippocampus. Lesion studies, such as the one conducted on the famous patient Henry Molaison (known in literature as H.M.), have shown that hippocampal lesions may have been responsible for his retrograde amnesia or his inability to remember new events. These were not unequivocal, however, as the surgery on H.M. removed not only the anterior portion of the hippocampus, but also much of the hippocampal gyrus and the amygdala.

The discovery that LTP contributes to memory consolidation is based on the premise that memories might be explained by changes in neural connections, just like learning. In the 1960s, researchers in Oslo observed long-term forms of synaptic plasticity in the hippocampus, yet they did not notice the possible link between plasticity and memory. This has paved the way, over the years, for further discoveries by new researchers, notably Bliss, who appeared in the lab in Oslo. In 1973, researchers published a paper, which was first to describe LTP (Bliss and Gardner-Medwin 1973). The paper hypothesized these forms of long-term synaptic plasticity could be the mechanism for memory consolidation, with the hippocampus being identified as having features required by LTP. On one hand, it has a psychological function (supporting memories), and on the other hand, its components have synaptic plasticity. It's a structure on the intermediate level between cellular neuroscience and psychology.

But without further research, the hypothesis that LTP in the hippocampus is responsible for memory would be just one of the myriad of other ideas proposed in the search for the engram. Different fields, however, have brought their own constraints on LTP:

> The findings in … varied fields and from these different perspectives added their own constraints on the mechanism of LTP. Different perspectives could explore, for example, different components of the mechanism, different properties of those components, different activities in which those components engage or different forms of organization among them (Craver 2003, 187).

But it is notable that the story does not end here. What researchers wanted to integrate next was a molecular-level mechanism for LTP. Bickle (2003, 62–75) reconstructs one of the molecular hypotheses, but fails to mention it is just one of the many, or too many, hypotheses of mechanisms responsible for a cellular-level behavior (Sanes and Lichtman 1999; Malenka and Bear 2004). One of the critical problems is that LTP is a physiological phenomenon, and our current experimental techniques are just too crude to intervene in it precisely. It's difficult to design an experiment that would (i) confirm LTP contributes to memory consolidation not only in the hippocampus and elsewhere and (ii) indicate its crucial molecular components.

Note that one of the implicitly accepted constraints is there is a single molecular component of synaptic plasticity involved in LTP and LTD (long-term depression of excitatory synaptic transmission). In their influential review paper, Malenka and Bear state that LTP and LTD are not unitary phenomena because “their mechanisms vary depending on the synapses and circuits in which they operate” (2004, 5). The research assumption is that the cellular phenomenon is unitary, only if it has the same molecular mechanism that does not vary depending on the surrounding context. Failing to constrain models offered in various experimental settings has led them, therefore, to deny LTP is one component; rather, it would rather be a family of components.

The constraint just mentioned is not a *finding* in itself. One could make two decisions faced with the evidence cited by Malenka and Bear. First, one could specify the phenomenon that the LTP is supposed to give rise to in more abstract terms, which would lead to making the question the model is supposed to answer a little more abstract. Or, second, one could retain the question and make multiple LTP models more specific. These strategies are called *lumping* and *splitting* (Craver 2009, 581–82). While these strategies may fail or succeed, depending on how the world is, they are not findings or discoveries. Lumping and splitting strategies may create idealized models; we could simply distort reality for tractability, simplicity, or some other reason. One important factor in the lumping strategy is it could also help theoretical unification, which might allow researchers to answer questions why some mechanisms work in a similar way by appealing to their similar causal structure (Weber and Lefevere 2015).

Interestingly, in the subsequent years, a new molecular candidate for the component of LTP was found, protein kinase Mzeta (PKMζ). The initial studies demonstrated a role for the kinase in memory maintenance; disrupting PKMζ activity with ζ-inhibitory peptide (ZIP) succeeded in disrupting many established associations in several key brain regions (which is a bottom-up intervention). More recent work, however, has questioned the role of PKMζ in memory maintenance and the effectiveness of ZIP as a specific inhibitor of PKMζ activity, but was not conclusive (for review, see Kwapis and Helmstetter 2014). It turns out, however, the reason ZIP is ineffective is there is a compensatory mechanism, activated if PKMζ should fail (Tsokas et al. 2016).

So, while causal interventions are difficult because of the complex organization of the memory consolidation mechanism, the current experimental evidence seems in favor of the lumping strategy. The complex organization is to be expected; after all, memory consolidation is one of the most important functions of the brain, and it may contain backup systems. This also shows simplistic molecular models—which ignore the cellular and psychological mechanisms—are not entirely appropriate because experimental interventions that disrupt memory in a physiologically unrealistic fashion are held to be inconclusive. Physiologically realistic conditions are important because our experimental techniques are often too fat-handed and disrupt more components than intended. Which conditions are physiologically realistic is determined also by our behavioral knowledge: this is why the most cited studies on LTP come from experiments in vivo in animals in single-trial learning (Whitlock 2006).

To sum up, multiple constraints are at work in the current research on LTP as the component of the mechanism of memory consolidation, occurring both in the hippocampus and outside it. These constraints are related to experimental findings and to strategies, such as lumping and splitting, which are important in creating integrated models. Just because there is (as it seems) a mechanism that satisfies all the constraints, it is possible to build a coherent model.

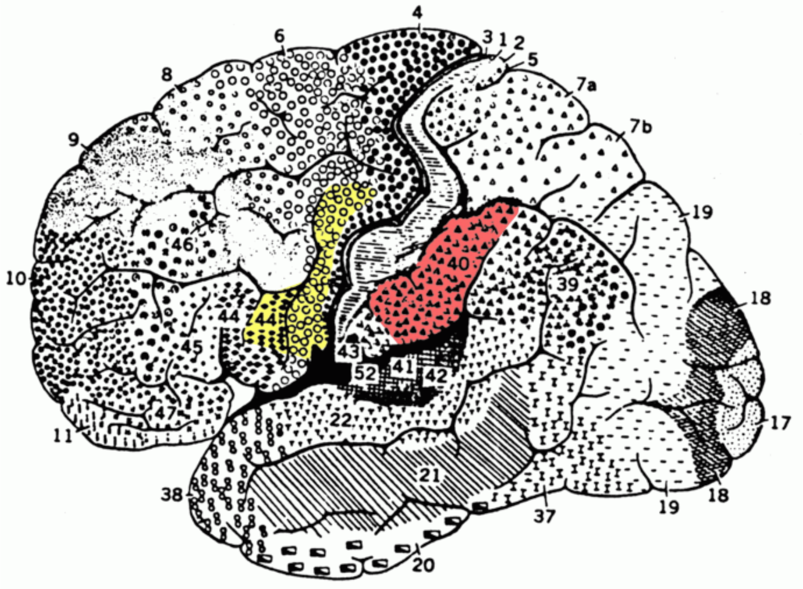
As Craver stresses, the research on LTP is highly interdisciplinary, combining electrophysiological and biochemical manipulations to explore the functioning of the proposed memory mechanism. This means it is not the case that it is *just* an inter-level integration of two models of mechanisms. In inter-level integration, the lower-level mechanism is a proper spatiotemporal part of the upper-level mechanism. Rather, multiple considerations are at play, and they mutually constrain hypotheses about models of LTP. Arguably, researchers treat models as constituting a distributed explanation of a mechanism (Hochstein 2015), and this seems to be the case with LTP. Researchers use many incomplete models of various properties, components, and operations mutually to constrain their overall explanation of the mechanism of LTP.

## 4. Failures of integration: mirror neurons and the gestural account of evolution of language

In this section, I will show how purported mechanistic integration can fail. To do so, I will show that mere consistency of mechanistic hypotheses is not enough to achieve integration. A case in point is the explanation of the mechanism responsible for sexual orientation in terms of mirror neurons (Ponseti et al. 2006). I will argue that, even if this account of sexual orientation is compatible with the existence of mirror neurons, the hypothesis that mirror neurons support action orientation does not constrain models of sexual orientation. The hypothesis that mirror neurons are responsible for action understanding does not furnish the mirror neuron account of sexual orientation with more explanatory depth.

In the 1990s, neuroscientists in Parma localized discharges of a group of neurons in both area F5 of the premotor cortex and in parietal area PF of the brain of macaques (di Pellegrino et al. 1992). Surprisingly, such discharges were reported both when the macaque performed an action and when it observed another individual performing a similar action. A similar fronto-parietial network, including the posterior inferior frontal gyrus, adjacent ventral premotor cortex, and the inferior parietal lobule was also observed in human brains (Rizzolatti and Craighero 2004), where the structural activations were observed in subjects observing and imitating actions (see Figure 2). This neural system responsible for action observation/execution matching was dubbed as the "mirror neuron system" (MNS). The MNS was hypothesized to be involved in diverse cognitive functions, including empathy (Gallese 2003), action understanding (Kohler et al. 2002), intention understanding, linguistic communication (Arbib 2005; Arbib 2012), and finally, sexual preferences (Ponseti et al. 2006; Mouras et al. 2008). This discovery, if real, would be crucial for further development of cognitive neuroscience and cognitive science (for a recent review, see (Kilner and Lemon 2013)).

Figure 2: Mirror neuron system (colored) in humans. Source: *Scholarpedia*.



Some theorists, however, are still critical. To begin, there were no direct recordings of mirror neurons in human beings, and there is only a functional homologue, whereas some fMRI studies are ambiguous (Lingnau, Gesierich, and Caramazza 2009). Others have criticized vehemently the use of the MNS hypothesis outside the domain of action imitation and understanding. Hickok (2014) also claims the lesions of the MNS do not lead to inhibition of action understanding.

The issue with the MNS goes deeper as there are multiple attempts at theoretical and mechanistic integration that fail because the MNS does not explain the operations of larger cognitive mechanisms. Selective discharges of the MNS neurons cannot explain complex dispositions, such as sexual preferences, as observing any other action leads to a selective discharge of such neurons. The mention of the MNS plays an ornamental role in sketchy boxological models that do not place any constraints on underlying mechanisms. There are virtually no constraints on intention understanding, unless we can tell what 'intentions' are (and one can doubt whether these are really helpful theoretical posits; see (Nanay 2013)).

Let me turn then to the account that seems to embody a failure of integration, the account of sexual orientation in terms of mirror neurons. In a brain imaging study, which consisted of showing the images of aroused human genitals—with no context—to exclusively homosexual or exclusively heterosexual participants, it was found that ventral premotor cortex, "which is a key structure for imitative (mirror neurons) and tool-related (canonical neurons) actions showed a bilateral sexual preference-specific activation, suggesting that viewing sexually aroused genitals of the preferred sex triggers action representations of sexual behavior" (Ponseti et al. 2006, 825). The hypothesis was that three regions of interest (ROIs) might be involved: two regions related to the reward system—Centromedian Thalamus (and adjacent ventral striatum) and Orbitofrontal Cortex—and Ventral Premotor Cortex—responsible for hypothesized „motor representations of sexual behavior."

Instead of speaking of a mechanism for sexual orientation, the authors chose the term *endophenotype*, which is used to talk of hereditary characteristics of a certain condition, which are not its direct symptom. Using the term is not limited to evolutionary biology, but also covers psychiatric syndromes. However, the term is not defined in the paper, but used in the phrase *functional endophenotype*: "we propose that the observed response pattern represents a functional endophenotype for sexual orientation in humans" (Ponseti et al. 2006, 832). It does not seem to be a huge stretch to think the activity of the mechanism, whose characteristics are hereditary, is supposed to be causally relevant for sexual arousal compatible with one's sexual orientation (a follow-up paper no longer uses the term and links mirror neurons to male sexual arousal; (cf. Mouras et al. 2008)). Observing aroused genitals is supposed to trigger action understanding, and the activation of the ROIs in the fMRI study is correlated with genitals being compatible with one's sexual orientation.

As Hickok (2014) points out, it is not controversial that human beings can understand actions and imitate them. However, there is barely any evidence for the MNS in humans, and the Broca's area (the homologue of F5 in macaque monkeys) does not seem to have properties of mirror neurons in macaques. Hickok points out that actions are understood by human beings, even when the Broca area is lesioned, which should be impossible. Hence, a dissociation suggests there is a lack of the mechanistic constraint, which means action understanding can proceed without the purported MNS. By looking at apraxias (motor disorders), Hickok also argues there is a dissociation between the system responsible for action understanding and action production.

Does the activation of MNS explain sexual orientation? Obviously, mere correlation cannot establish the causal connection, and no lesion studies with ROIs observed during the experiments are cited in the paper. No interventions using transcranial magnetic stimulation (TMS) were done either. So, the evidence supports only the hypothesis that the regions are possible mechanisms for sexual orientation. The hypothesis is also incompatible with previous findings about mirror neurons: they are supposed to fire when the animal observes a specific action executed by another animal and when the animal performs it. So, it would seem they should fire when one observes genitals compatible with one's own sex, irrelevant of one's sexual orientation. These are compatible with one's actions, in contrast to the genitals of a sexual partner. This is not even discussed in the paper (which therefore violates this mechanistic top-down constraint), but let's suppose there may be an explanation why this is not the case.

The paper fails also to be constrained by our previous knowledge about sexual arousal (violates componency constraints). Even if sexual orientation has an affectional component, there is evidence that sexual desire and romantic love can be dissociated (Diamond 2003); the two can be studied separately. But it would be more appropriate to say the study was concerned with sexual arousal, and hypothesized one factor of the arousal was sexual orientation. So, it is natural to ask whether the ROIs under study were the same as previously studied for, say, male sexual arousal. Previous Positron Emission Tomography (PET) studies have shown the highest activation in the claustrum, a region whose function had been unclear. Activations were recorded in paralimbic areas (anterior cingulate gyrus, orbitofrontal cortex), in the striatum (head of caudate nucleus, putamen), and in the posterior hypothalamus (Redouté et al. 2000). But the a priori hypotheses in the fMRI studies were apparently not wholly positively constrained by this study when picking the ROIs; particularly, the claustrum was not studied. Instead, they were constrained by the mirror-neuron sensorimotor hypothesis.

Let me wrap up. The study of the functioning sexual arousal mechanism, cryptically dubbed as *functional endophenotype*, as influenced by sexual orientation, is a clear example of a failure to constrain the models of mechanisms of sexual arousal both by previous studies on the topic (using PET scans) and by the methodological constraints that require more than mere correlation. Instead, mere correlation was observed in the hypothesized system for "motor representations of sexual behavior", localized in the Ventral Premotor Cortex.

The hypotheses do not yield a well-supported model of the actual mechanism of sexual arousal and do not explain why the arousal is linked to pictures of genitals of the preferred sex, rather than the pictures of one's own sex. All these failures are failures to constrain the hypothetical model of the "functional endophenotype" or, rather, a mechanism of sexual arousal as linked with a certain sexual orientation.

## 5. Conclusion

In this paper, I claimed the mechanistic integration of models can be viewed in terms of constraints on scientific representations of mechanisms. Both the success of integration and its failure are linked to stronger or weaker constraints on multiple models of mechanisms. With LTP, numerous mechanistic constraints on multiple models of mechanisms seem to create an inter-field research framework on the phenomenon, which is driven by constraints on possible unitary models of memory consolidation (the lumping strategy is the default). With the sexual arousal in relation to MNS, there are almost no mechanistic constraints from MNS. The models of sexual arousal are not even truth-constrained by previous arousal models, and merely consistent with MNS models.

I have argued that constraints are best understood in terms of certain representations—be it experimental findings, theoretical principles, or methodological norms—acting upon possible representations of mechanisms. This modification of the notion of constraint, as used previously in the mechanistic approaches to integration, has several advantages. First, some constraints used by Craver and Darden are not findings—at least not experimental findings. These constraints may also include methodological norms and strategies, such as lumping or splitting. Second, there is a clear connection to the account on constraint-satisfaction developed by Thagard, so integration of models can proceed semi-automatically. Third, properly constrained mechanistic models may be idealized to be explanatorily appropriate, and to understand this, one should allow that constraints operate on representations, rather than act directly on the space of possible mechanisms.

Without mechanistic constraints, there is no integration of mechanistic models. The future work on constraints of mechanistic models should analyze multiple kinds of constraints in greater detail and discover different constraints recognized by modelers. This could be helpful in distinguishing different kinds of integration, which would further develop Craver and Darden's work. The claim it is the web of constraints that underlies integration seems sufficiently well-established to believe that such progress is possible.

Acknowledgements: The work on this paper was funded from National Science Centre research grant under the decision DEC-2014/14/E/HS1/00803. The author wishes to thank three anonymous reviewers and Ken Aizawa, Joe Dewhurst, Paweł Grabarczyk, Piotr Makowski and Alfredo Vernazzani for comments, as well as Rafał Czajkowski for his helpful remarks and literature suggestions related to the contemporary research on LTP. All faults remaining in the paper are of course only mine.

## References

Arbib, M. A. 2005. From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics. *Behavioral and Brain Sciences* 28 (2): 105-24-67. doi:10.1017/S0140525X05000038.

Arbib, M. A. 2012. *How the Brain Got Language: The Mirror System Hypothesis*. New York: Oxford University Press.

Bechtel, W. 1986. The Nature of Scientific Integration. In *Integrating Scientific Disciplines*, W. Bechtel ed., 3–52. Dordrecht: Springer Netherlands. doi:10.1007/978-94-010-9435-1_1.

Bechtel, W., Abrahamsen, A. 2005. Explanation: A Mechanist Alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2): 421–41. doi:10.1016/j.shpsc.2005.03.010.

Bickle, J. 2003. *Philosophy and Neuroscience*. Dordrecht: Springer Netherlands. doi:10.1007/978-94-010-0237-0.

Bliss, T. V. P., Gardner-Medwin, A. R. 1973. Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Unanaestetized Rabbit Following Stimulation of the Perforant Path. *The Journal of Physiology* 232 (2): 357–74.

Bliss, T. V P, Gardner-Medwin, A. R., Lømo, T. 1973. Synaptic Plasticity in the Hippocampal Formation. In *Macromolecules and Behaviour,* edited by G. B. Ansell and P. B. Bradley, 193–203. London: Macmillan.

Boone, W., Piccinini, G. 2015. The Cognitive Neuroscience Revolution. *Synthese*, June. Springer Netherlands. doi:10.1007/s11229-015-0783-4.

Busemeyer, J. R., Diederich, A. 2010. *Cognitive Modeling*. BOOK. Los Angeles: Sage.

Craver, C. F. 2002. Interlevel Experiments and Multilevel Mechanisms in the Neuroscience of Memory. *Philosophy of Science* 69 (S3). UChicago Press: 83–97.

Craver, C. F. 2003. The Making of a Memory Mechanism. *Journal of the History of Biology* 36 (1): 153–95. doi:10.1023/A:1022596107834.

Craver, C. F. 2005. Beyond Reduction: Mechanisms, Multifield Integration and the Unity of Neuroscience. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36 (2): 373–95. doi:10.1016/j.shpsc.2005.03.008.

Craver, C. F. 2007. *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, C. F. 2009. Mechanisms and Natural Kinds. *Philosophical Psychology* 22 (5): 575–94. doi:10.1080/09515080903238930.

Craver, C. F., Darden, L. 2013. *In Search of Mechanisms: Discoveries across the Life Sciences*.

Cummins, R. 1975. Functional Analysis. *The Journal of Philosophy* 72 (20): 741–65.

Danks, D. 2014. *Unifying the Mind: Cognitive Representations as Graphical Models.* Cambridge, Mass.: MIT Press.

Darden, L., Maull N. 1977. Interfield Theories. *Philosophy of Science* 44 (1): 43–64.

Diamond, L. M. 2003. What Does Sexual Orientation Orient? A Biobehavioral Model Distinguishing Romantic Love and Sexual Desire. *Psychological Review* 110 (1): 173–92. doi:10.1037/0033-295X.110.1.173.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti G. 1992. Understanding Motor Events: A Neurophysiological Study. *Experimental Brain Research* 91 (1): 176–80.

Gallese, V. 2003. The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity. *Psychopathology* 36 (4): 171–80. doi:10.1159/000072786.

Glennan, S. S. 2002. Rethinking Mechanistic Explanation. *Philosophy of Science* 69 (S3): S342–53. doi:10.1086/341857.

Glymour, C. N. 1994. On the Methods of Cognitive Neuropsychology. *The British Journal for the Philosophy of Science* 45 (3): 815–35. doi:10.1093/bjps/45.3.815.

Glymour, C. N. 2001. *The Mind's Arrows. Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, Mass.: MIT Press.

Green, J. D., Adey W. R. 1956. Electrophysiological Studies of Hippocampal Connections and Excitability. *Electroencephalography and Clinical Neurophysiology* 8 (2): 245–63. doi:http://dx.doi.org/10.1016/0013-4694(56)90117-1.

Hensel, W. M. 2013. On Reduction and Interfield Integration in Neuroscience. In *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental,* edited by Marcin Miłkowski and Konrad Talmont-Kamiński, 167–81. Newcastle upon Tyne: Cambridge Scholars Publishing.

Hickok, G. 2014. *The Myth of Mirror Neurons: The Real Neuroscience of Communication and Cognition*. New York: WW Norton.

Hochstein, E. 2015. One Mechanism, Many Models: A Distributed Theory of Mechanistic Explanation. *Synthese*. Springer Netherlands. doi:10.1007/s11229-015-0844-8.

Illari, M. P., Williamson J. 2011. What Is a Mechanism? Thinking about Mechanisms across the Sciences. *European Journal for Philosophy of Science* 2 (1): 119–35. doi:10.1007/s13194-011-0038-2.

Irvine, E. 2014. Models, Robustness, and Non-Causal Explanation: A Foray into Cognitive Science and Biology. *Synthese*, July, n/a-n/a. doi:10.1007/s11229-014-0524-0.

Kilner, J. M., Lemon R.N. 2013. What We Know Currently about Mirror Neurons. *Current Biology* 23 (23): R1057–62. doi:10.1016/j.cub.2013.10.051.

Kohler, E., Keysers Ch., Umiltà M. A., Fogassi L., Gallese V., Giacomo R. 2002. Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons. *Science (New York, N.Y.)* 297 (5582): 846–48. doi:10.1126/science.1070311.

Kwapis, J. L., Helmstetter F. J. 2014. Does PKM(zeta) Maintain Memory? *Brain Research Bulletin* 105 (June). NIH Public Access: 36–45. doi:10.1016/j.brainresbull.2013.09.005.

Lashley, K. S. 1950. In Search of the Engram. In *Physiological Mechanisms in Animal Behavior. (Society's Symposium IV.).*, 454–82. Oxford: Academic Press, Society for Experimental Biology.

Levy, A., Bechtel W. 2013. Abstraction and the Organization of Mechanisms. *Philosophy of Science* 80 (2): 241–61. doi:10.1086/670300.

Lingnau, A., Gesierich B., Caramazza A. 2009. Asymmetric fMRI Adaptation Reveals No Evidence for Mirror Neurons in Humans. *Proceedings of the National Academy of Sciences of the United States of America* 106 (24): 9925–30. doi:10.1073/pnas.0902262106.

Machamer, P., Darden L., Craver C. F. 2000. Thinking about Mechanisms. *Philosophy of Science* 67 (1): 1–25.

Malenka, R.C., Bear M. F. 2004. LTP and LTD. *Neuron* 44 (1): 5–21. doi:10.1016/j.neuron.2004.09.012.

Miłkowski, M. 2013. *Explaining the Computational Mind*. Cambridge, Mass.: MIT Press.

Miłkowski, M. 2016. Explanatory Completeness and Idealization in Large Brain Simulations: A Mechanistic Perspective. *Synthese* 193 (5): 1457–78. doi:10.1007/s11229-015-0731-3.

Miłkowski, Marcin. 2017. Unification Strategies in Cognitive Science. *Studies in Logic, Grammar and Rhetoric* 48, no. 1: 13–33. doi:10.1515/slgr-2016-0053.

Mouras, H., Stoléru S., Moulier V., Pélégrini-Issac M., Rouxel R., Grandjean B., Glutron, D., Bittoun, J. 2008. Activation of Mirror-Neuron System by Erotic Video Clips Predicts Degree of Induced Erection: An fMRI Study. *NeuroImage* 42: 1142–50. doi:10.1016/j.neuroimage.2008.05.051.

Nagel, E. 1961. *The Structure of Science Problems in the Logic of Scientific Explanation.* New York: Harcourt Brace & World.

Nanay, B. 2013. *Between Perception and Action*. Oxford: Oxford University Press.

Newell, A. 1973. You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. In *Visual Information Processing*, edited by W. G. Chase, 283–308. New York: Academic Press.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, Mass. and London: Harvard University Press.

Piccinini, G., Craver C. F. 2011. Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches. *Synthese* 183 (3): 283–311. doi:10.1007/s11229-011-9898-4.

Ponseti, J., Bosinski H. A., Wolff S., Peller M., Jansen O., Mehdorn H. M., Büchel Ch., Siebner H. R. 2006. A Functional Endophenotype for Sexual Orientation in Humans. *NeuroImage* 33 (3): 825–33. doi:10.1016/j.neuroimage.2006.08.002.

Redouté, J., Stoléru S., Grégoire, M. C., Costes, N., Cinotti L., Lavenne F., Le Bars D., Forest M. G., Pujol J.F. 2000. Brain Processing of Visinottiual Sexual Stimuli in Human Males. *Human Brain Mapping* 11 (3). John Wiley & Sons, Inc.: 162–77. doi:10.1002/1097-0193(200011)11:3<162::AID-HBM30>3.0.CO;2-A.

Rizzolatti, G., Craighero L. 2004. The Mirror-Neuron System. *Annual Review of Neuroscience* 27 (January): 169–92. doi:10.1146/annurev.neuro.27.070203.144230.

Sanes, J. R., Lichtman J. W. 1999. Can Molecules Explain Long-Term Potentiation? *Nature Neuroscience* 2 (7): 597–604. doi:10.1038/10154.

Spirtes, P., Glymour C. N., Scheines R. 2000. *Causation, Prediction, and Search*. 2nded. Cambridge, Mass.: The MIT Press.

Thagard, P. 2000. *Coherence in Thought and Action*. Cambridge, Mass.: MIT Press.

Thagard, P. 2007. Coherence, Truth, and the Development of Scientific Knowledge. *Philosophy of Science* 74: 28–47.

Triantafillou, S., Tsamardinos, I. 2015. Constraint-Based Causal Discovery from Multiple Interventions over Overlapping Variable Sets. *Journal of Machine Learning Research* 16 (March): 2147–2205.

Tsokas, P., Hsieh Ch., Yao Y., Lesburguères, E., Wallace E. J. C., Tcherepanov A., Jothianandan D., et al. 2016. Compensation for PKMζ in Long-Term Potentiation and Spatial Long-Term Memory in Mutant Mice. *eLife* 5 (May). doi:10.7554/eLife.14846.

Van Orden, G., Pennington B. F., Stone G. O. 2001. What Do Double Dissociations Prove? *Cognitive Science* 25 (June): 111–72.

Weber, E., Lefevere M. 2015. Unification, the Answer to Resemblance Questions. *Synthese*. Springer Netherlands. doi:10.1007/s11229-015-0969-9.

Whitlock, J. R. 2006. Learning Induces Long-Term Potentiation in the Hippocampus. *Science* 313 (5790): 1093–97. doi:10.1126/science.1128134.