DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 48 (61) 2016

DOI: 10.1515/slgr-2016-0053

Marcin Miłkowski
Institute of Philosophy and Sociology
Polish Academy of Sciences

UNIFICATION STRATEGIES IN COGNITIVE SCIENCE*

Abstract. Cognitive science is an interdisciplinary conglomerate of various research fields and disciplines, which increases the risk of fragmentation of cognitive theories. However, while most previous work has focused on theoretical integration, some kinds of integration may turn out to be monstrous, or result in superficially lumped and unrelated bodies of knowledge. In this paper, I distinguish theoretical integration from theoretical unification, and propose some analyses of theoretical unification dimensions. Moreover, two research strategies that are supposed to lead to unification are analyzed in terms of the mechanistic account of explanation. Finally, I argue that theoretical unification is not an absolute requirement from the mechanistic perspective, and that strategies aiming at unification may be premature in fields where there are multiple conflicting explanatory models.

Keywords: cognitive science, unification, integration, simplicity, invariance, monstrosity.

1. Whence unification?

The need for unified models, theories or conceptual frameworks in cognitive science may seem self-explanatory from today's point of view. However, this need was not at all obvious earlier in cognitive psychology. In 1973, Allen Newell, commenting on papers submitted to a conference on visual processing, argued that cognitive psychology can no longer remain fragmented (Newell 1973). The state of affairs was, according to him, quite detrimental to the prospects of developing a general account of cognition. Psychology dealt with individual phenomena – such as the continuous rotation effect; chess position perception; linear search on displays; the serial position effect in free recall; perceptual illusions; ambiguous figures; or the visual icon – that were studied in separate tasks given to experimental subjects. To conceptualize them theoretically, psychologists referred to binary oppositions, such as *nature versus nurture*; *serial versus parallel processing*;

ISBN 978-83-7431-512-8 ISSN 0860-150X

analog versus digital; *conscious versus unconscious*; *stages versus continuous development; motor versus pure perception in perceptual learning*; and so forth. But progress could not be expected:

> Suppose that in the next thirty years we continued as we are now going. Another hundred phenomena, give or take a few dozen, will have been discovered and explored. Another forty oppositions will have been posited and their resolution initiated. Will psychology then have come of age? Will it provide the kind of encompassing of its subject matter – the behavior of man – that we all posit as a characteristic of a mature science? And if so, how will the transformation be accomplished by this succession of phenomena and oppositions? (Newell 1973, pp. 287–288)

The oppositions were too crude to serve as pointers to a general theory, and phenomena were too detailed to furnish researchers with a framework that could be confirmed empirically. The oppositions were becoming less and less clear, and no unity in explanations was to be found.

Instead, Newell proposed that one should focus on what he called *a unified theory of cognition*. His proposal, as is well known, was to study *integrated* cognitive architectures designed to perform all the individual tasks studied by cognitive psychologists (Newell 1990). Unified cognitive architectures would alleviate the worries voiced so prominently by Newell: Simply because there would be a single, highly structured entity capable of information-processing in all these tasks, there would be a unified account of the reason that various tasks are actually performed.

His proposal is interesting because it does not follow the traditional recipe for unification as defended by philosophers of science, i.e., it does not advocate theory reduction, at least not in its classical version. In essence, the classical account of theory reduction follows the logic of the received view of explanation as defended by Hempel and Oppenheim (1948): to explain is to present a sound deductive argument. In the case of explanation and prediction, the premises describe laws of science and antecedent conditions, while the conclusion states the description of the phenomenon to be explained.

Similarly, in the case of theory reduction, the premises contain the laws of the new theory, T_1 and bridge laws that connect the terms of theory T_2 with terms of theory T_1, while the conclusion describes T_2. While this formulation is not without flaws (for refined formulations, see Hooker 1981a; Hooker 1981b; Hooker 1981c; Churchland 1985; Schaffner 1993; Bickle 1998), it is certainly quite elegant, not least because of its relative simplicity. The important feature of this logical account of reduction is supposed to make unification and integration of theories inextricable. I will argue below that

integration and unification of scientific representations should be distinguished, and that some kinds of integration may lead to disunion. That may be disadvantageous, so there are reasons to defend the classical account despite its problems.

This classical account remains largely impractical for one very simple reason: non-fundamental sciences usually do not contain any laws in the classical Hempelian sense, i.e., universally quantified, true statements of *unlimited scope* without designations of any particular objects. Most biological regularities, even if analyzed as law-like *ceteris paribus* statements, are results of quite particular historical and evolutionary circumstances. They are essentially of limited scope and usually limited (even if implicitly) to the results of evolutionary processes. Moreover, psychology does not seem to feature even limited regularities (Cummins 2000); it normally describes individual events, phenomena, and their mechanisms.

It is the lack of laws in psychology and other cognitive science disciplines that justifies the adoption of an alternative account of explanation. In this paper, I will embrace the mechanistic account of explanation (Craver 2007; Bechtel 2008; Miłkowski 2013). According to this account, to explain a phenomenon is to describe the underlying mechanism responsible for it. Mechanisms are organized systems, composed of entities and activities (also called interactions, operations, or simply processes by various authors). Their overall causal structure gives rise to one or more phenomena to be explained. In some important respects, mechanistic explanation assumes principles of functional analysis as advocated by proponents of functionalism (Cummins 1984; Fodor 1968), but requires the components and activities to be causally relevant for the phenomenon as well (Piccinini & Craver 2011).

The new mechanistic approach to explanation is sometimes presented as non-reductive, but this characterization is confusing and misleading as it does not – in contrast to functionalism – advocate for autonomy of special sciences at all (Hensel 2013; Boone & Piccinini 2015). On the other hand, and in an important sense, mechanistic constitutive explanations are reductive: they explain how some phenomena occur in terms of component entities and activities of mechanisms, which are located at lower levels of organization (in a mechanistic understanding of the term: see (Craver 2007)) rather than in mechanisms themselves. Constitutive explanations are said to provide *deeper* understanding of phenomena (Thagard 2007) because they expose the causal structure that stands behind the phenomena to be explained. This mechanistic kind of reduction is not classical, but it justifies certain explanatory identities. Hence, some mechanists have explicitly

avowed identity theory as an important research heuristic (Bechtel & McCauley 1999). Quite obviously, the mechanistic reduction is not eliminative; rather, it substantiates the claim that the higher levels of mechanistic organization exist as compositions of lower levels of entities and activities.

2. Integration versus unification

The terms *integration* and *unification* are sometimes used interchangeably and without further explication. In this paper, I will distinguish them in the following way: Explanatory unification is the process of developing general, simple, elegant, and beautiful explanations, while explanatory integration is the process of combining multiple explanations in a coherent manner. Similarly, one can also define methodological unification as the process of developing general-purpose, simple research methods; and methodological integration as the process of combining multiple methods in research.

Classical reduction was supposed to deliver theories that were both explanatorily integrated and unified. Things are not so simple, though. 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 without positing any substantial connections between its claims and the claims of T_1 (cf. Bechtel, 1986, p. 41). In general, most methods of integration and unification do not guarantee that both occur at the same time. However, many defenders of mechanistic explanation conflate the issues of integration and unification. For example, Piccinini and Craver write: "we sketch a framework for building a unified science of cognition. This unification is achieved by showing how functional analyses of cognitive capacities can be and in some cases have been integrated with the multilevel mechanistic explanations of neural systems" (Piccinini & Craver, 2011, p. 284). In this paper, I argue for distinguishing both issues more carefully.

In general, defenders of mechanistic explanation are particularly sensitive to issues of integration (Bechtel 1986; Craver & Darden, 2013) and to inter-field research (Darden & Maull, 1977), which relates at least two fields of study. By a "field of study", Darden and Maull understand "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."[1] (Darden & Maull, 1977, p. 44) Two fields

of study may appeal to the same or overlapping spatiotemporal locations, entities, or activities, and one of them may provide a better understanding of the spatiotemporal relationships, causal relationships, physical nature, structure, or function thereof. In the case of cognition, it is quite clear that cognitive processes may be explained in various ways by various disciplines.

Cognitive science is composed of multiple fields with stronger and weaker connections among them. The stronger the connections between fields A and B, the bigger the chance that models will integrate insights from A and B. In previous studies of mechanistic integration, at least three modes of integration of mechanisms, and therefore also fields, were identified: *Simple integration*, when the models of mechanisms can be considered as pieces of a puzzle that fit together; *inter-level relationship*, when another level of organization is added to make explanation more complete; and *inter-temporal integration* (Craver & Darden, 2013, Chapter 10). In the case of simple integration, two fields may simply study cognition in a similar way but with a slightly varying emphasis on each mechanism. Notice that in this case, both models are at the same level of organization, so simple integration is not inter-level. The inter-level integration usually involves deepening an existing explanation of a phenomenon by providing an underlying mechanism of the phenomenon, as in the case of providing the non-classical, mechanistic reduction introduced in this section. These three patterns of integration can be easily observed in cognitive science. However, their list is neither a systematic taxonomy of all possible ways that mechanisms can be integrated, nor does it provide a deep understanding of integration as such. The observed patterns of integration correspond to the spatial and temporal adjacency (simple integration) and spatial or temporal containment (inter-level and inter-temporal integration).

Craver proposes to understand integration in terms of constraints on the space of possible mechanisms. A constraint is "a finding that either shapes the boundaries of the space of plausible mechanisms or changes the probability distribution over that space" (Craver, 2007, p. 247). However, there are two reasons to modify his account: First, some theoretical or methodological principles may drive the search for plausible mechanisms in the space, and principles can only metaphorically be understood as findings. Second, genuinely satisfying explanations of mechanisms may involve idealization. Idealized mechanisms may be physically or even logically impossible, as they are often introduced as intentional simplifications or distortions, hence they cannot be found in the space of plausible mechanisms. For this reason, I will account for the search for adequate mechanistic explanations in the space of plausible representations of mechanisms.

The notion of constraint is therefore understood in terms of a representation that shapes the boundaries of the space of plausible representations of mechanisms or the probability distribution over that space. To make it more precise, one may also integrate this account of mechanistic constraints with another recent general account of inter-theoretic and inter-model relationships. According to Danks, "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, p. 31) This means that if S constrains T because of a certain theoretical virtue, then if we care about this virtue in T, we should care about it in S. Differing virtues give rise to differing kinds of constraints.

The weakest kind of constraint is a *truth-constraint*: two bodies of knowledge satisfy a truth-constraint in case they can be both true at the same time. However, truth-constraining is a weak relation of logical coherence. Note that attaining coherence – by satisfying constraints – between various representations (and models of various kinds) was studied by Thagard in his account of integration, too (Thagard 2000). Integrating possibly incoherent scientific representations is therefore a kind of *coherence problem*, which is defined in the following 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, p. 18).

He notices, however, that coherence problems are, in general, NP-complete, or practically intractable (Thagard, 2000, p. 28). In other words, an algorithm based on simple exhaustive search will not be tractable for coherence problems, unless the case is trivially simple (i.e., contains a small number of elements). Instead, he offers several heuristic search strategies that approximate a satisfactory solution. However, because models in cognitive science, even if stated in a machine-readable form, are rarely integrated automatically (especially if they are supposed to conform to semantic con-

straints that refer to the spatial and temporal properties of entities and activities in mechanisms), integrating them remains more art than science. The constraint-satisfaction account of integration for mechanistic models does not serve merely a practical purpose. The constraint satisfaction account can describe all previously found kinds of integration, and more, so it is a slightly more general proposal for a unifying account of mechanistic integration.

It is notable, however, that the results of integration need not be simple, beautiful, or general. These properties are notoriously difficult to pin down precisely, but the idea here is very simple: Even if multiple constraints are in operation, the resulting scientific representation may be highly redundant, violate parsimony considerations, and so forth. Even if mechanistic constraints are preserved, the resultant representation may be quite disconnected; for example, one can integrate the account of the cognitive map in the hippocampus (Derdikman & Moser, 2010) with, say, Baddeley's account of working memory (Baddeley & Hitch, 1974). Both models refer to working memory but as Baddeley (2000) notes, they use the notion to mean different things; hence, even if rats have both kinds of memory, no explanatory unity is observed here. So the next question to consider is: What is explanatory unity?

3. Dimensions of unification

Intuitively, unified explanations are simple, general, and beautiful. The appeal to aesthetic criteria may seem to invoke non-analyzable, elusive properties, and perhaps this is the reason why unification has rarely been analyzed by the defenders of mechanistic explanation. There are, however, at least three properties ascribed to unified explanations:

1. invariance or unbounded scope;
2. simplicity or lack of redundancy;
3. elegance or beauty.

Let us consider these properties in turn. The unbounded scope of explanation is sometimes held to be its necessary feature. For example, Hempel and Oppenheim required laws of nature to be of unbounded scope simply because law-based explanations are supposed to have unlimited scope. Similarly, in this tradition, Philip Kitcher has defended his account of explanatory unification by appealing to the large scope of unified explanations (Kitcher 1989). The opponents of this account point out that explanatory power does not depend on the number of phenomena to be explained. For example, a theory

that explains the Big Bang does not seem to be less explanatorily powerful than a model that explains two car accidents in Warsaw, even if, nominally, the first one, has smaller scope. Yet the Big Bang is of much greater scientific significance. Admittedly, however, even if defenders of causal explanation do not require such explanations to be necessarily of unlimited scope, they would agree that good explanatory models should have (practically) maximal possible scope. At the same time, mechanists require these models to be causal: "unification is genuinely explanatory when it refers to higher-level structure of common mechanisms" (Glennan, 2002, p. S352).

What about simplicity? The classical principle of ontological parsimony is that entities should not be multiplied beyond necessity, which is simply Occam's Razor. In contemporary terms, ontological parsimony involves the number of ontological commitments of a theory (Quine 1948). However, one might want to go beyond mere ontological parsimony to assess the simplicity of theories. There are multiple ways to analyze this notion. For example, one account is due to Popper who claimed that the simpler theory is the one that is more falsifiable (Popper 1959). This simple account is however open to many objections, which show that it is counterintuitive. As Nelson Goodman (1961) noted, the hypothesis "All maple trees are deciduous", is intuitively simpler than the hypothesis, "All maple trees whatsoever, and all sassafras trees in Eagleville, are deciduous", but the latter is more falsifiable.

Still, there were some efforts to define the measures of simplicity more formally, and they were usually not prone to simple counterexamples. It's arguable that simplicity may be analyzed in statistical terms (for example, Akaike information criterion (see Forster & Sober, 1994)). Such criteria can be quite easily applied to analyze computational models in cognitive science (Busemeyer & Diederich, 2010). Another way to analyze formal models is to appeal to algorithmic information theory, which defines Kolmogorov-Solomonoff complexity (Chaitin 1987; Li & Vitanyi, 1993). In the latter case, the redundancy of scientific representation may be approximated simply in terms of the compressed model's size, and the model is compressed by some general-purpose algorithm of lossless compression (such as PPM, or prediction by partial matching, or ZIP; for such empirical investigation, see (Zenil 2010)).

While it may be argued that various measures of parsimony or simplicity yield different notions, all of them show that simplicity is not just a result of intuitive judgment, and further work is required to see how it is connected to unification.

Elegance or beauty may seem the most difficult to analyze. After all, notions such as *beauty* are primarily aesthetic, and it may be controversial

to judge which theory is beautiful and which not. However, the problem has been recently approached from another angle by Ioannis Votsis (2015) who proposed to start from the opposite end: from monstrosity. He analyzes the notion of monstrosity in terms of the lack of shared relevant deductive consequences. Intuitively, a theory is monstrous only if it contains "isolated islands" that are confirmationally disconnected, i.e., what these "islands" imply is completely disjoint. To spell this notion out more precisely, Votsis refers to the notion of relevant deductive inference defined by Schurz: "a valid deduction is relevant iff no subformula of the conclusion is replaceable on some of its occurrences by any other formula *salva validitate* of the deduction" (Schurz, 1991, p. 391). The shorthand notation for "y is a *relevant deductive consequence* of x" is '$x \vdash_r y$'. Using this notion, Votsis defines *confirmational disconnectedness* thus:

> Any two content parts of a non-self-contradictory proposition Γ expressed as propositions A, B are confirmationally disconnected if, and only if, for all pairs of internally non-superfluous propositions α, β where $A \vdash_r \alpha$ and $B \vdash_r \beta$: (i) there is no true or partly true proposition γ such that $\alpha \vdash_r \gamma$ and $\beta \vdash_r \gamma$ and (ii) where $0 < P(\alpha), P(\beta) < 1$, $P(\alpha/\beta) = P(\alpha)$ and (iii) there is no atomic proposition δ such that $\alpha \wedge \beta \vdash_r \delta$, $\alpha \vdash_r \delta$ and $\beta \vdash_r \delta$ (Votsis, 2015, p. 102)

While this formulation may, again, be impractical for (partially) informal theories in cognitive science, as well as for non-verbal computational models, which are not stated as interpreted logical calculi at all, it offers a valuable explication of the notion of monstrosity. One may think of more practical ways of assessing monstrosity, for example in terms of the statistical independence of two parts, A and B, of a scientific representation: $P(A \cap B) = P(A)P(B)$, which sometimes may be estimated in terms of the mutual information of simulation models (for performance criteria of simulation models, see Hora & Campos, 2015). However, a review of possible methods for assessing confirmational disconnectedness statistically goes beyond the scope of this paper.

Let me wrap up this section. Unified scientific representations may be analyzed in terms of (a) their unbounded scope; (b) simplicity; (c) lack of monstrosity. These features are maximized by a simple statement of one universal law that is true of everything and whose formulation has no component parts. Indeed, this may be the intuition behind the search for a grand theory of everything: it would be maximally unified if its statement were extremely simple and universal. However, these features have all been defined as measures on some scale, which means that scientific representations may be assessed as more or less unified. It is important to see that these features

are not wholly interdependent: a representation may be true of just one thing and remain maximally simple and non-monstrous. Similarly, a non-monstrous representation that contains several parts may not be maximally simple (as redundancy does not increase monstrosity). But maximally simple representation (say, expressed as a single propositional variable p) may not be monstrous, so these dimensions of unification are not totally independent either. This raises the open question of whether there are more dimensions in the unification of scientific representations.

4. Two popular strategies

The practice of unifying models or theories in cognitive science does not simply boil down to an application of unification criteria or even benchmarks. This is because representations to be unified are not yet even stated (completely). Instead, researchers adopt unification strategies, two of which have been identified by David Danks (2014). These can be analyzed in line with the mechanistic account of explanation, as will be shown below. This analysis will show that from the mechanistic perspective, these are actually three individual unification strategies.

The first strategy appeals to schemes of structures: "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" (Danks, 2014, p. 176). Quite clearly, Newell's use of cognitive architectures to unify theories of cognition fits into this category. Cognitive architectures are systems that can perform multiple cognitive tasks using the same structure, which makes the explanation invariant in this respect: the internal structure stays constant regardless of the external environment, and makes the explanation more unified, or parsimonious. Cognitive architectures have remained immensely influential, so multiple schemes of structures are used to unify theories of cognition (for a recent review, see Byrne 2012). This includes both traditional architectures such as SOAR (Laird, Newell & Rosenbloom, 1987), which has its roots in Newell's research, ACT-R (Anderson 2007), which remains influential in psychology, and architectures that strive for neuro-scientific plausibility such as Leabra (O'Reilly & Munakata, 2000) and SPAUN (Eliasmith et al., 2012). Another example of this kind is research on cognitive robotic architectures as unifying cognition (Morse et al., 2011). Thus, to study developmental processes, one may use one of the robotic platforms of so-called epigenetic robotics, such as i-Cub (Metta et al., 2010).

The second strategy is an appeal to elementary processes. Researchers strive to show how "coherent cognition arises from shared processes, where those processes are typically small building blocks that combine to yield complex cognition" (Danks, 2014, p. 177). Note that this is not the kind of piecemeal approach criticized by Newell: the elementary building blocks and their interactions are supposed to be at play in multiple individual tasks. This approach has its roots deeply in the Cartesian proposal to understand the work of the nervous system in terms of the reflex arc: the stimulus pulls tiny wires of the nervous system, which in turn open little valves in the brain, releasing animal spirits to hollow the nerve tubes that lead to appropriate muscles. The nervous system is simply a collection of reflex arcs under this approach. A similar approach can also be found in computational modeling of the nervous system: already in the first model (McCulloch & Pitts, 1943), it was proposed that coherent cognition is the product of the complex interaction of neurons understood as logical gates or computational devices that embody logical operators such as conjunction (AND gate) and disjunction (OR gate).

The contemporary connectionist modeling research program adopts the same strategy: the nervous system is composed of a number of similar computational units, which are more or less biologically plausible. Note that this approach can mesh easily with unified theories of cognition as long as the pattern of connections between these units is not simply fitted to observed data but results from theoretical considerations. Such is the case with SPAUN, which is essentially a connectionist neural network composed of spiking neurons. However, their connections are not trained using machine-learning algorithms; instead, they are set up according to hypotheses about the function of certain brain areas (Eliasmith 2013). The same applies to a contemporary proposal for a unified theory of cognition in terms of predictive coding (Clark, 2016, 2013): there is a certain high-level functional pattern of the whole cognitive system, which is said to implement strategies that approximate Bayesian reasoning in perception and action, and the function is implemented by a hierarchy of similar units that perform predictive coding and send error information to other levels of the hierarchy.

In other words, both strategies can be complementary, and do not exclude one another. It's worthwhile analyzing them also in mechanistic terms. The first strategy is straightforward: there is a *mechanism schema*, or an incomplete representation of entities and activities interacting together, which contains gaps to be filled (Craver 2007). Because these gaps are sometimes filled just to fit the observational data, critics argue that cognitive architectures have limited explanatory power (Roberts & Pashler, 2000). From

the mechanistic point of view this criticism is justified to some extent: the explanatory power is not just a matter of the fit between the mechanistic model and the data, but of the adequacy of relevant causal hypotheses. Hence, the model's accuracy requires not just the fit between the performance of psychological subjects but also the performing of bottom-up and top-down interventions in the mechanism (Craver 2002).

The second strategy turns out to have two different versions: one may posit the same mechanism for various phenomena, or multiple similar mechanisms for similar phenomena.

The first case is easily illustrated with mirror neurons. 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 macaque brains (di Pellegrino et al., 1992). 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, the adjacent ventral premotor cortex, and the inferior parietal lobule, was also observed in human brains (Rizzolatti & Craighero, 2004), where the structural activations were observed in subjects observing and imitating actions. This neural system responsible for action observation/execution matching was called the *mirror neuron system* (MNS). The MNS was hypothesized to be involved in quite diverse cognitive functions, including empathy (Gallese 2003), action understanding (Kohler et al., 2002), intention understanding, linguistic communication (Arbib 2005; Arbib 2012), and even sexual preferences (Ponseti et al., 2006; Mouras et al., 2008). However, as it turns out, some such hypotheses are based on spurious correlations, and top-down interventions are ignored in this research. For example, as Hickok (2014) argues, if the MNS is responsible for action understanding, a lesion of the MNS should lead to a deficit in action understanding. But it does not, and experiments demonstrating such are simply ignored.

Even more problematic is that the overall structure of the mechanism remains largely sketchy. A *mechanism sketch* is a representation of a mechanism that lacks its crucial entities and activities; it does not even contain placeholder terms (Craver 2007). Explanations that appeal to a mechanism sketch are not (entirely) successful. The problem is that it also remains unclear how the MNS is supposed, for example, to influence sexual orientation *exactly*. Mere selective discharge of this area is not sufficient to establish its causal relevance. Extrapolation of the MNS to explain ever new domains of cognition often remains speculative, as long as there are no independent causal interventions that could deliver new empirical evidence. This,

of course, is not to say that a single elementary mechanism may not be involved in multiple phenomena. But for all phenomena, the causal relevance should be established independently.

The second case, of positing multiple similar mechanisms for similar phenomena, may be illustrated with neurons posited as individual components of the nervous system by Santiago Ramon y Cajal (Ramón y Cajal 1990). Quite clearly, there are many kinds of neurons but their operations are in some important respects similar (with some notable exceptions such as "silent neurons").

Therefore, the application of unification strategies for elementary structures is based on extrapolation, or transposition of the same model of mechanism to ever new explananda. As one reduces the number of individual explanations, the redundancy is limited, and simplicity increases, which in turn is one of the major features of unification. The same reason makes the first strategy, of hypothesizing the same overall mechanism structure for various phenomena, unificatory. At the same time, with multiple explanatory hypotheses bound to the same mechanism, one keeps monstrosity at bay. Instead of yet another mechanism for every explanandum, one may appeal to the same one or at least to the same type of mechanism. The same goes for scope: we explain multiple phenomena with the (type of) mechanism M, so the explanatory scope of M increases with each phenomenon.

It remains to be discussed whether unification is a norm of mechanistic explanation, or just a non-mandatory practice. I will argue for the latter claim.

5. Unification versus mechanistic norms of explanation in cognitive science

Proponents of the new mechanistic philosophy have not underlined the value of unification as much as they have embraced integration. The importance of explanatory unification has been emphasized by proponents of frameworks that were supposed to be an alternative to causal explanation; this is how Philip Kitcher has framed his proposal, by opposing Wesley Salmon's account of causal explanation (Salmon 1998). For example, he claimed that:

> The heart of the unification approach is that we cannot make sense of the notion of a basic mechanism apart from a systemization of the world in which as many consequences as possible are traced to the action of as small a number of basic mechanisms as possible (Kitcher, 1989, p. 497).

But how would one justify this claim? After all, it is quite plausible that there may be a large number of basic mechanisms out there. Stuart Glennan (2002, p. S352) offers the following reply to Kitcher's argument: Even if our heuristics of search and discovery of mechanisms are biased towards the discovery of a small number thereof, these heuristics may fail, and we can easily make sense of the (possible) world in which a plethora of various basic mechanisms is at work. Similarly, the simplicity and scope of our theories is a convenient assumption that may be easily dismissed as soon as we discover, for example, that biological mechanisms are not optimally but only sufficiently simple to remain reliable: there is a clear trade-off between simplicity of design and redundancy.

This means that the striving for explanatory unification, in contrast to integration, is not be an absolute norm for defenders of the mechanistic account of explanation. Instead, unification should be considered to be an epistemological virtue of scientific representations rather than of mechanisms described by these representations. But it is not mandatory for explanations to be genuine or satisfying. To show this more clearly, let me discuss three approaches to unification: simplicity, invariance and unbounded scope, and non-monstrosity.

Obviously, simpler and non-redundant representations are preferred, as long as they remain tractable or useful for our representational purposes. This point has avoided the attention of proponents of parsimony and simplicity in the past: maximally non-redundant representations may be difficult to decipher. Let's take a simple example, one of the simplest axiomatizations of the propositional calculus, offered by Jan Łukasiewicz in his notation: *EEpqEErqEpr*. The notation is obscure even to those trained in Reverse Polish Notation. Similarly, a plain text compressed by a general-purpose algorithm is no longer human-readable. Removing redundancy comes at a cost: first, it may make the representation more susceptible to error (as redundancy helps error detection); second, it requires more computational effort to handle non-redundant representation. For this reason, models of mechanisms should be as simple and parsimonious *only* as far as it aids their uses.

Models of mechanisms that describe more invariant causal structures are also useful to the point where they still remain tractable or readable on pain of Bonini's paradox: the model may be as difficult to understand as the phenomenon under modeling, and for complex artificial networks simulating the brain, the paradox looms large (Dawson, 1998, p. 17). However, it does not seem to be a norm of mechanistic explanation that they have unbounded scope. Some biological regularities may occur only in cer-

tain spatiotemporal locations, and causal explanations seem mostly local. This does not make them any less explanatorily powerful. Similarly, an explanation that addresses a single phenomenon is not necessarily worse than one that explains two phenomena. What is more important is how significant these phenomena are. Their significance may be assessed, for example, in terms of consequences for other scientific representations of the world. To use Quine's metaphor of the web of belief to describe the scientific representations as connected together, the representations at the periphery are probably less significant, while the ones closer to the center are more germane to others. This metaphorical picture should be sufficient for our purposes here; it is obvious that models of mechanisms that are more significant should be valued over ones at the periphery unless there is some reason to believe that there may be a large uncharted territory ahead, and that a given model is just the beginning of a successful research paradigm.

Similarly, non-monstrosity is to be preferred but only when there is reason to believe that maximizing confirmational connectedness preserves truth. Why? Because structures may exist that are composed of relatively independent subsystems, and a model that would describe these subsystems as totally interdependent would be at best an idealization, and at worst, wishful thinking.

Therefore, as my discussion indicates, no feature of unification mentioned above is an absolute ideal for the mechanistic account of explanation. One could reply that this is because explanatory models are in some way special, i.e., they need not be unified to be satisfying, whereas there are some models, in particular physical ones, that stand in need of genuine unification. This is the kind of argument that was put forward by proponents of robotic architectures of cognition (Morse et al., 2011): To make a cognitive robot work, one needs a unified and complete model of its cognitive capacities. But this argument is not valid. While one needs to build a physically complete robot, it does not mean that all its features need to be completely modeled in a theoretically unified fashion in order for it to work. Quite the contrary, some details of the physical implementation may remain unknown before one starts to actually build physical models; quick and dirty tricks may be enough to make them work. Moreover, the existence of hybrid robotic models that link together quite diverse approaches to cognitive and motor capacities in the same physical entity shows that unification is not strictly required to make such robots. All that is really required is not unification but simplicity; invariance or non-monstrosity are not at all necessary. This seems to suggest that even for models that

are not just explanatory, but for example, exploratory, unification is not an absolute norm.

To repeat, constraints that force unification on models are just heuristic biases that may help to develop beautiful models. However, confusing these heuristics with infallible rules may lead to detrimental consequences. Indeed, in the past, the principle of parsimony was abused. Even if it says that *entia non sunt multiplicanda praeter necessitatem*, or that entities should not be multiplied beyond necessity, the last two words – beyond necessity – seem to have been ignored by zealots of parsimony. Therefore, behaviorists would deny the existence of entities that were inconsistent with their theories rather than ones that were redundant, to mention only cognitive maps, still being debated in the 1990s (Benhamou 1996), even if the opposite hypothesis was experimentally idle for further research on rat spatial navigation (Bechtel 2016). The same applies to consciousness that had to be investigated in the U.S. under the term *attention* to eschew the behaviorist dogma.

Simplifying the view of the world beyond necessity leads to dogmatism. One should not deny our ignorance of the world and its complex phenomena. Complex phenomena are difficult to explain, and conflicting models thereof may be useful in several ways. First, multiple contradictory and idealized models may be built to derive inferences about robust regularities in operations within a given system (Weisberg 2006). This is how current climate models operate, and we may envisage that brain simulations could be built in a similar fashion. Second, contradictions between models is fuel for progress in developing further models. No model ever explains in a theoretical void; models explain in a distributed fashion (Hochstein 2015). But distributed explanations should not be contradictory, so one needs to build coherent representations, and in doing so, monstrous explanations should be avoided if possible.

"Strive for unified explanations!" is therefore just a useful heuristic but not an absolute norm. It might indeed turn out that cognitive systems are collections of semi-independent cognitive modules, as defenders of evolutionary psychology and massive modularity have claimed (Cosmides & Tooby, 1987; but see Richardson 2007 for a mechanistic criticism). But before assuming a priori that cognitive systems are unified or not, we should first try to see how experimental evidence may affect the issue, and this opens really difficult questions (Van Orden & Kloos, 2003). In general, complex models cannot be easily falsified or fitted to data, and their usefulness may be assessed only in terms of Lakatosian progressive or regressive research programs (Cooper 2007). As things stand right now, both approaches

seem to be similarly fruitful. A mechanist should therefore applaud and let a thousand flowers bloom. Picking the flowers comes later.

NOTES

* The work on this paper was funded by a National Science Centre (Poland) research grant under the decision DEC-2014/14/E/HS1/00803. The author wishes to thank Daniel Kostic, Michał Klincewicz, Hubert Kowalewski, Ricardo Sanz and the audience during the 11th Congress of the Polish Society for Cognitive Science for comments on a previous version of this paper.

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

REFERENCES

Anderson, J. R. (2007). *How Can the Mind Occur in the Physical Universe?* Oxford: Oxford University Press.

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. https://doi.org/10.1017/S0140525X05000038

Arbib, M. A. (2012). *How the brain got language: the mirror system hypothesis.* New York: Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). New York / San Francisco / London: Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

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

Bechtel, W. (2008). *Mental Mechanisms.* New York: Routledge (Taylor & Francis Group).

Bechtel, W. (2016). Investigating neural representations: the tale of place cells. *Synthese*, *193*(5), 1287–1321. https://doi.org/10.1007/s11229-014-0480-8

Bechtel, W., & McCauley, R. N. (1999). Heuristic identity theory (or back to the future): The mind-body problem against the background of research strategies in cognitive neuroscience. In *Proceedings of the 21st Annual Meeting of the Cognitive Science Society* (pp. 67–72). Mahwah, NJ: Erlbaum.

Benhamou, S. (1996). No evidence for cognitive mapping in rats. *Animal Behaviour*, *52*(1), 201–212.

Bickle, J. (1998). *Psychoneural reduction the new wave*. Cambridge, Mass.: MIT Press.

Boone, W., & Piccinini, G. (2015). The cognitive neuroscience revolution. *Synthese*. https://doi.org/10.1007/s11229-015-0783-4

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Los Angeles: Sage.

Byrne, M. D. (2012). Unified theories of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(4), 431–438. https://doi.org/10.1002/wcs.1180

Chaitin, G. J. (1987). *Algorithmic information theory*. Cambridge [Cambridgeshire]; New York: Cambridge University Press.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *The Journal of Philosophy*, *82*(1), 8–28.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2016). *Surfing uncertainty: prediction, action, and the embodied mind*.

Cooper, R. P. (2007). The Role of Falsification in the Development of Cognitive Architectures: Insights from a Lakatosian Analysis. *Cognitive Science*, *31*(3), 509–533. https://doi.org/10.1080/15326900701326592

Cosmides, L., & Tooby, J. (1987). From Evolution to Behavior: Evolutionary Psychology as the Missing Link. In J. Dupre (Ed.), *The Latest on the Best. Essays on Evolution and Optimality* (pp. 277–303). Cambridge, Mass.: MIT Press.

Craver, C. F. (2002). Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philosophy of Science*, *69*(S3), 83–97.

Craver, C. F. (2007a). *Explaining the Brain. Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, C. F. (2007b). *Explaining the Brain. Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: discoveries across the life sciences*.

Cummins, R. (1984). Functional analysis. *Conceptual Issues in Evolutionary Biology: An Anthology*.

Cummins, R. (2000). "How does it work" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. A. Wilson (Eds.), *Explanation and Cognition* (pp. 117–145). Cambridge, Mass.: MIT Press.

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.

Dawson, M. (1998). *Understanding cognitive science*. Malden Mass.: Blackwell.

Derdikman, D., & Moser, E. I. (2010). A manifold of spatial maps in the brain. *Trends in Cognitive Sciences*, *14*(12), 561–9. https://doi.org/10.1016/j.tics.2010.09.004

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.

Eliasmith, C. (2013). *How to build the brain: a neural architecture for biological cognition*. New York: Oxford University Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., ... Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*, *338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266

Fodor, J. A. (1968). *Psychological explanation: an introduction to the philosophy of psychology*. New York: Random House.

Forster, M., & Sober, E. (1994). How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1–35. https://doi.org/10.1093/bjps/45.1.1

Gallese, V. (2003). The Roots of Empathy: The Shared Manifold Hypothesis and the Neural Basis of Intersubjectivity. *Psychopathology*, *36*(4), 171–180. https://doi.org/10.1159/000072786

Glennan, S. S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, *69*(S3), S342–S353. https://doi.org/10.1086/341857

Goodman, N. (1961). Safety, Strength, Simplicity. *Philosophy of Science*, *28*(2), 150–151. https://doi.org/10.1086/287795

Hempel, C., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.

Hensel, W. M. (2013). On Reduction and Interfield Integration in Neuroscience. In M. Miłkowski & K. Talmont-Kamiński (Eds.), *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental* (pp. 167–181). 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*. https://doi.org/10.1007/s11229-015-0844-8

Hooker, C. A. (1981a). Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. *Dialogue*, *20*(1), 38–59. https://doi.org/10.1017/S0012217300023088

Hooker, C. A. (1981b). Towards a General Theory of Reduction. Part II: Identity in Reduction. *Dialogue*, *20*(2), 201–236. https://doi.org/10.1017/S0012217300023301

Hooker, C. A. (1981c). Towards a General Theory of Reduction. Part III: Cross-Categorical Reduction. *Dialogue*, *20*(3), 496–529. https://doi.org/10.1017/S0012217300023593

Hora, J., & Campos, P. (2015). A review of performance criteria to validate simulation models. *Expert Systems*, *32*(5), 578–595. https://doi.org/10.1111/exsy.12111

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher & W. C. Salmon (Eds.), *Scientific Explanation* (Vol. 505, pp. 410–505). Minneapolis: University of Minnesota Press.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science (New York, N.Y.)*, *297*(5582), 846–8. https://doi.org/10.1126/science.1070311

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1–64. https://doi.org/10.1016/0004-3702(87)90050-6

Li, M., & Vitanyi, P. (1993). *An Introduction to Kolmogorov Complexity and Its Applications*. New York, Berlin, Heidelberg: Springer-Verlag.

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., ... Montesano, L. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, *23*(8), 1125–1134. https://doi.org/10.1016/j.neunet.2010.08.010

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

Morse, A. F., Herrera, C., Clowes, R., Montebelli, A., & Ziemke, T. (2011). The role of robotic modelling in cognitive science. *New Ideas in Psychology*, *29*(3), 312–324. https://doi.org/10.1016/j.newideapsych.2011.02.001

Mouras, H., Stoléru, S., Moulier, V., Pélégrini-Issac, M., Rouxel, R., Grandjean, B., ... Bittoun, J. (2008). Activation of mirror-neuron system by erotic video clips predicts degree of induced erection: an fMRI study. *NeuroImage*, *42*, 1142–1150. https://doi.org/10.1016/j.neuroimage.2008.05.051

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

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

O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience. Understanding the Mind by Simulating the Brain*. Cambridge, Mass.: MIT Press.

Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4

Ponseti, J., Bosinski, H. A., Wolff, S., Peller, M., Jansen, O., Mehdorn, H. M., ... Siebner, H. R. (2006). A functional endophenotype for sexual orientation in humans. *NeuroImage*, *33*(3), 825–833. https://doi.org/10.1016/j.neuroimage.2006.08.002

Popper, K. R. (1959). *The logic of scientific discovery. New Yorker, The*. Hutchinson.

Quine, W. V. (1948). On What There Is. *The Review of Metaphysics*, *2*(5), 21–38.

Ramón y Cajal, S. (1990). *New ideas on the structure of the nervous system in man and vertebrates*. (L. W. Swanson, Trans.). Cambridge, Mass.: MIT Press.

Richardson, R. C. (2007). *Evolutionary Psychology as Maladapted Psychology*. Cambridge, Mass.: MIT Press.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–92. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358.

Salmon, W. C. (1998). *Causality and explanation*. Oxford University Press, USA.

Schaffner, K. F. (1993). *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press.

Schurz, G. (1991). Relevant Deduction: From Solving Paradoxes Towards a General Theory. *Erkenntnis*, *35*, 391–437. https://doi.org/10.1007/BF00388295

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.

Van Orden, G. C., & Kloos, H. (2003). The Module Mistake. *Cortex*, *39*(1), 164–166. https://doi.org/10.1016/S0010-9452(08)70092-3

Votsis, I. (2015). Unification: Not Just a Thing of Beauty. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *30*(1), 97. https://doi.org/10.1387/theoria.12695

Weisberg, M. (2006). Robustness Analysis. *Philosophy of Science*, *73*(5), 730–742. https://doi.org/10.1086/518628

Zenil, H. (2010). Compression-based investigation of the dynamical properties of cellular automata and other systems. *Journal of Complex Systems*, *19*(1).