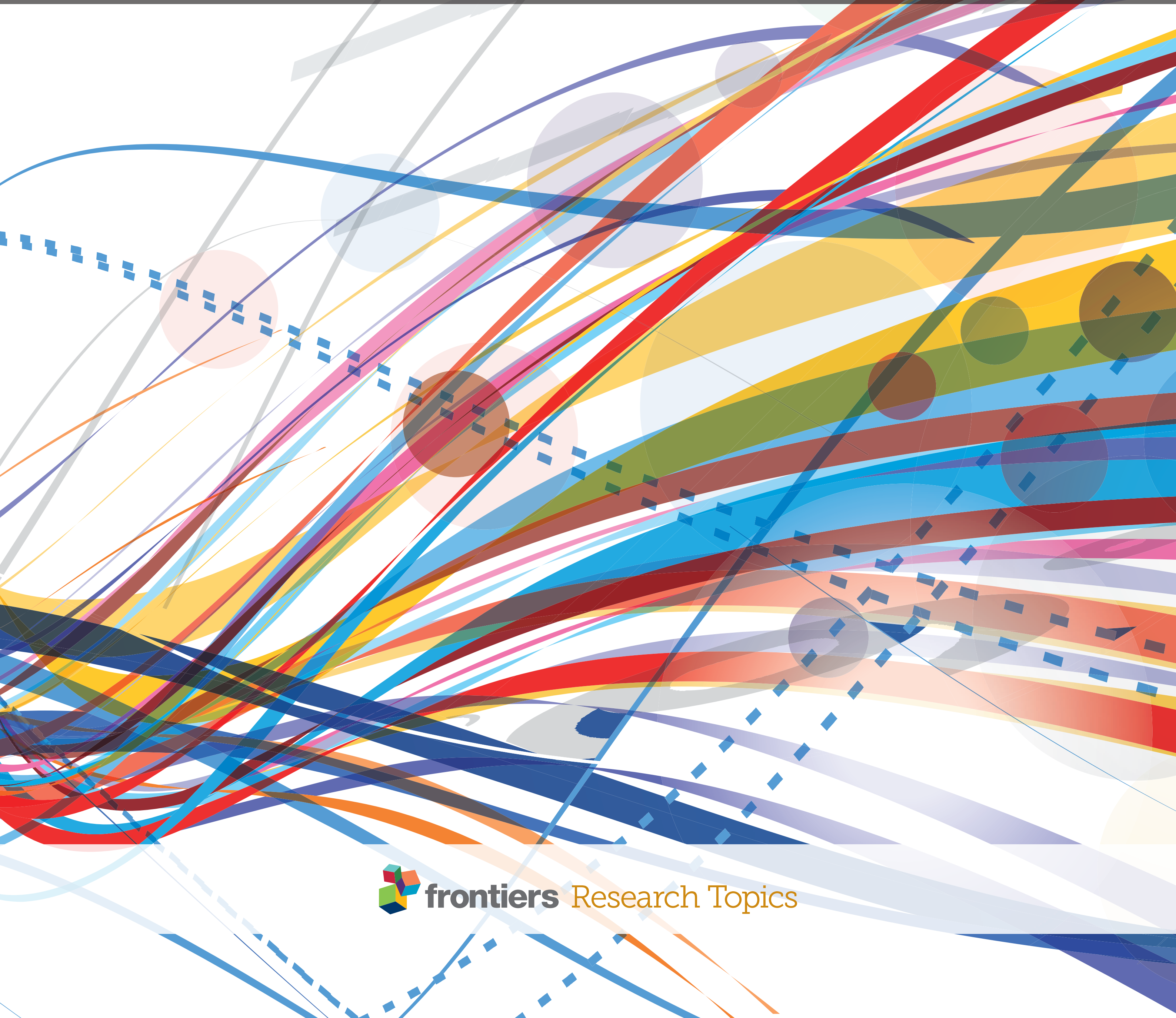
WHAT LEVELS OF EXPLANATION IN THE BEHAVIOURAL SCIENCES?

EDITED BY: Giuseppe Boccignone and Roberto Cordeschi

PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88919-597-8
DOI 10.3389/978-2-88919-597-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

WHAT LEVELS OF EXPLANATION IN THE BEHAVIOURAL SCIENCES?

Topic Editors:
Giuseppe Boccignone, University of Milan, Italy
Roberto Cordeschi, Sapienza University of Rome, Italy

Complex systems are to be seen as typically having multiple levels of organization. For instance, in the behavioural and cognitive sciences, there has been a long lasting trend, promoted by the seminal work of David Marr, putting focus on three distinct levels of analysis: the computational level, accounting for the What and Why issues, the algorithmic and the implementational levels specifying the How problem.

However, the tremendous developments in neuroscience knowledge about processes at different scales of organization together with the complexity of today cognitive theories suggest that there will hardly be only three levels of explanation. Instead, there will be many different degrees of commitments corresponding to the different granularities—from high-level (behavioural) models to low-level (neural and molecular) models of the cognitive research program. For instance, Bayesian approaches, that are usually advocated for formalizing Marr's computational level and rational behaviour, have even been adopted to model synaptic plasticity and axon guidance by molecular gradients. As a result, we can consider the behavioural scientist as dealing with models at a multiplicity of levels.

The purpose of this Research Topic in Frontiers in Theoretical and Philosophical Psychology is to promote an approach to the role of the levels and explanation and models which is of interest for cognitive scientists, neuroscientists, psychologists, behavioural scientists, and philosophers of science.

Citation: Boccignone, G., Cordeschi, R., eds. (2015). What Levels of Explanation in the Behavioural Sciences? Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-597-8

Table of Contents

frontiers in PSYCHOLOGY

EDITORIAL
published: 25 February 2015
doi: 10.3389/fpsyg.2015.00213

Coping with levels of explanation in the behavioral sciences

Giuseppe Boccignone[1]* and Roberto Cordeschi[2]

[1] Department of Computer Science, University of Milan, Milan, Italy
[2] Department of Philosophy, Sapienza University of Rome, Rome, Italy
*Correspondence: giuseppe.boccignone@unimi.it

Edited and reviewed by:
Axel Cleeremans, Université Libre de Bruxelles, Belgium

Keywords: human behavior, levels of explanation, neuroscientific models, functional models, computational models, cognitive architectures, mechanisms, reductionism

INTRODUCTION

This Research Topic aimed at deepening our understanding of the levels and explanations that are of interest for cognitive scientists, neuroscientists, psychologists, behavioral scientists, and philosophers of science.

Indeed, contemporary developments in neuroscience and psychology suggest that scientists are likely to deal with a multiplicity of levels, where each of the different levels entails laws of behavior appropriate to that level (Berntson et al., 2012). Also, gathering and modeling data at the different levels of analysis is not sufficient: the integration of information across levels of analysis is a crucial issue.

Given such state of affairs, a number of interesting questions arise. How can the autonomy of explanatory levels be properly understood in behavioral explanation? Is reductionism a satisfactory strategy? How can high-level and low-level models be constrained in order to be actually explanatory of both behavioral and neurological or molecular evidence? What is the kind of relationship between those models?

PLURALITY OF LEVELS WITH AND BEYOND MARR

Marr (1982) distinguished between three levels of explanation, the *what/why* level (computational theory), the *how* level (algorithm), and the *physical realization* level (implementation). His influential framework has had a far-reaching influence in both neuroscience and cognitive science over the years and it has become a sort of paradigm. However, the tremendous developments in such sciences suggest that there will hardly be only three levels of explanation.

For instance, Castelfranchi (2014) claims for several different layers of "theory": the cognitive representations and mechanisms; the neural processes; the evolutionary history and adaptive functions of our cognition and behaviors; the social structures and dynamics with their relation and feedbacks on individual minds and behaviors; the historical and cultural mechanisms; the developmental paths.

Clearly, on the one hand, dealing with such complexities calls for models that simulate those processes so that they can be used as explanatory tools, i.e., instances of the "synthetic" method (Cordeschi, 2002). In this perspective, Conte and Paolucci (2014) make the point that simple recipes have prevailed up to now and shadowed the application of rich cognitive models. As a viable solution, they discuss Agent Based Modeling and its role at the highest behavioral level of Computational Social Science.

On the other hand, to cope with multi-level complexity, Abney et al. (2014) propose explanatory pluralism. They present one concrete example, the analysis of a corpus of conversing individuals solving a joint decision-making task, performed by using decision-making at the behavioral level, confidence sharing at the linguistic level, acoustic energy at the physical level.

A further interesting issue is that of the objective vs. subjective meaning of the explanatory levels. Varma (2014) discusses how Marr's approach focused on the objective meaning of each level—how it supports computational models that correspond to cognitive phenomena—and he develops a complementary analysis of the subjective meaning of each level—how it helps cognitive scientists understand cognition. With the goal of showing that different kinds of explanation arise because we have different kinds of explanatory concerns, a clear case study is proposed by Wilkinson (2014) by using contrasting theories of delusional misidentification.

RELATIONSHIPS, CONSTRAINTS AND MECHANISMS

Addressing any level of description involves a certain degree of realist commitment at that level, which, in turn, has some consequences on the problems of reduction and of causality between levels. In this respect, one important case study is presented by Albertazzi and Poli (2014), who address the conundrum of color. They claim that color is a different entity for each level of reality and it generates different observables in the epistemologies of the different sciences.

Ramos (2014) introduces the hypothesis that the sophisticated psychological constructs classically associated with the concept of mental representation are essentially of the same nature of simple interactive behaviors. Thus, the capacity of generating elaborated mental phenomena like beliefs and desires emerges gradually during evolution, and social interaction. Here, mental representations are biological phenomena whose construction is achieved by a correlational mechanism of information exchange with the external world.

In a related perspective and in order to cope with the multiscale nature (Abney et al., 2014) of cognitive and behavioral phenomena, Costa and Ferraro (2014) argue that a statistical mechanics approach is almost inescapable. Starting from very simple systems, connectivity gives rise to levels of increasing functional complexity.

Here the key issue is that, at any level, systems obey laws holding for the lower levels; meanwhile, they are subjected to new constraints (related to and implemented through neural structures). These, in turn, generate new features, like novel patterns of activity, requiring adequate levels of representation in terms of model structures and variables.

Indeed, accounting for constraints is a central point: as Abney et al. (2014) put it, "mapping across levels should create mutual constraints, in that levels should be consistent, if qualitatively, with each other." A hallmark of the present state of research in cognitive/behavioral sciences is that one is generally ignorant of how exactly to cast the different levels into a grounded relationship. In this case the notions of structure and architecture—and related graphical modeling tools—become crucial, since they are necessary to embody constraints at the chosen level of explanation (Boccignone and Cordeschi, 2007, 2012).

The exploitation of structure/architecture as a tool for bridging intra- and inter-level constraints has the merit of paving the way for reconciling rational or information-based analyzes (Danks, 2008) with mechanism-based explanations (Bechtel and Abrahamsen, 2005). As fostered by Castelfranchi (2014), "laws" are not enough, both the "why" and "how" must be addressed.

In this respect, Datteri and Laudisa (2014) lucidly address the subtelties of graphical explanations, making the case for the relationship between box-and-arrow (BA) explanations and neuroscientific mechanism descriptions (NMDs). The interesting point raised by Datteri and Laudisa is that the BA analysis imposes constraints on the formulation of the NMD by postulating a number of regularities to be sought for in the neural activities of the system. Conversely, the NMD constrains the space of the possible BA analyzes of the system by postulating a number of neural regularities.

CONCLUSION

Taken together, the papers in "What levels of explanation in the behavioral sciences" give us some important indications of where the field is going and also demonstrate how lively and open the field is today.

We hope this Research Topic paves the way to new avenues and challenges for future work.

ACKNOWLEDGMENTS

Prof. Roberto Cordeschi prematurely passed away while we were in the process of evaluating the overall editorial result of this Research Topic. I have tried to ensure the accuracy of reporting what we had discussed so far, which is briefly summarized in this Editorial. Errors and omissions are, of course, my responsibility (GB).

REFERENCES

Abney, D. H., Dale, R., Yoshimi, J., Kello, C. T., Tylén, K., and Fusaroli, R. (2014). Joint perceptual decision-making: a case study in explanatory pluralism. *Front. Psychol.* 5:330. doi: 10.3389/fpsyg.2014.00330

Albertazzi, L., and Poli, R. (2014). Multi-leveled objects: color as a case study. *Front. Psychol.* 5:592. doi: 10.3389/fpsyg.2014.00592

Bechtel, W., and Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Stud. Hist. Phil. Biol. Biomed. Sci.* 36, 421–441. doi: 10.1016/j.shpsc.2005.03.010

Berntson, G. G., Norman, G. J., Hawkley, L. C., and Cacioppo, J. T. (2012). Evolution of neuroarchitecture, multi-level analyses and calibrative reductionism. *Interface Focus* 2, 65–73 doi: 10.1098/rsfs.2011.0063

Boccignone, G., and Cordeschi, R. (2007). "Bayesian models and simulations in cognitive science," in *Models and Simulations 2* (Tilburg, NL: PhilSci-Archive). Available online at: http://philsci-archive.pitt.edu/view/confandvol/2007005.html

Boccignone, G., and Cordeschi, R. (2012). Predictive brains: forethought and the levels of explanation. *Front. Psychol.* 3:511. doi: 10.3389/fpsyg.2012.00511

Castelfranchi, C. (2014). For a science of layered mechanisms: beyond laws, statistics, and correlations. *Front. Psychol.* 5:536. doi: 10.3389/fpsyg.2014.00536

Conte, R., and Paolucci, M. (2014). On agent-based modeling and computational social science. *Front. Psychol.* 5:668. doi: 10.3389/fpsyg.2014.00668

Cordeschi, R. (2002). *The Discovery of the Artificial: Behavior, Mind and Machines Before and Beyond Cybernetics*. Dordrecht: Kluwer.

Costa, T., and Ferraro, M. (2014). A statistical mechanical problem? *Front. Psychol.* 5:947. doi: 10.3389/fpsyg.2014.00947

Danks, D. (2008). "Rational analyses, instrumentalism, and implementations," in *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*, ed SM. Oaksford and N. Chater (Oxford: Oxford University Press), 59–75.

Datteri, E., and Laudisa, F. (2014). Box-and-arrow explanations need not be more abstract than neuroscientific mechanism descriptions. *Front. Psychol.* 5:464. doi: 10.3389/fpsyg.2014.00464

Marr, D. (1982). *Vision*. San Francisco, CA: W.H. Freeman.

Ramos, R. T. (2014). The concepts of representation and information in explanatory theories of human behavior. *Front. Psychol.* 5:1034. doi: 10.3389/fpsyg.2014.01034

Varma, S. (2014). The subjective meaning of cognitive architecture: a marrian analysis. *Front. Psychol.* 5:440. doi: 10.3389/fpsyg.2014.00440

Wilkinson, S. (2014). Levels and kinds of explanation: lessons from neuropsychiatry. *Front. Psychol.* 5:373. doi: 10.3389/fpsyg.2014.00373

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 January 2015; accepted: 11 February 2015; published online: 25 February 2015.

Citation: Boccignone G and Cordeschi R (2015) Coping with levels of explanation in the behavioral sciences. Front. Psychol. ***6****:213. doi: 10.3389/fpsyg.2015.00213*

This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 23 April 2014
doi: 10.3389/fpsyg.2014.00330

Joint perceptual decision-making: a case study in explanatory pluralism

Drew H. Abney[1]*, Rick Dale[1], Jeff Yoshimi[1], Chris T. Kello[1], Kristian Tylén[2,3] and Riccardo Fusaroli[2,3]

[1] Cognitive and Information Sciences, School of Social Sciences, Humanities and Arts, University of California, Merced, CA, USA
[2] Center for Semiotics, Aarhus University, Aarhus, Denmark
[3] Interacting Minds Center, Aarhus University, Aarhus, Denmark

Edited by:
Giuseppe Boccignone, Università di Milano, Italy

Reviewed by:
Felipe De Brigard, Duke University, USA
Russ McBride, University of Utah, USA

***Correspondence:**
Drew H. Abney, Cognitive and Information Sciences, School of Social Sciences, Humanities and Arts, University of California, 5200 North Lake Road, Merced, CA 95343, USA
e-mail: drewabney@gmail.com

Traditionally different approaches to the study of cognition have been viewed as competing explanatory frameworks. An alternative view, explanatory pluralism, regards different approaches to the study of cognition as complementary ways of studying the same phenomenon, at specific temporal and spatial scales, using appropriate methodological tools. Explanatory pluralism has been often described abstractly, but has rarely been applied to concrete cases. We present a case study of explanatory pluralism. We discuss three separate ways of studying the same phenomenon: a perceptual decision-making task (Bahrami et al., 2010), where pairs of subjects share information to jointly individuate an oddball stimulus among a set of distractors. Each approach analyzed the same corpus but targeted different units of analysis at different levels of description: decision-making at the behavioral level, confidence sharing at the linguistic level, and acoustic energy at the physical level. We discuss the utility of explanatory pluralism for describing this complex, multiscale phenomenon, show ways in which this case study sheds new light on the concept of pluralism, and highlight good practices to critically assess and complement approaches.

Keywords: explanatory pluralism, philosophy of science, joint decision-making, alignment, complexity matching

INTRODUCTION

Behavioral and cognitive processes are complex phenomena spanning multiple scales of organization, which may require multiple approaches to be fully understood. However, researchers have often aimed for a singular, unifying paradigm in the study of cognition (e.g., Fodor, 1975; Port and Van Gelder, 1995). The "paradigm wars" in cognitive science originated in the notion that one, or perhaps a limited number, of theoretical accounts will turn out to be most appropriate for the study of cognition. Herein, we will argue that multiple approaches should be used to study cognition at different scales of analysis. We consider a specific case study in detail, and show how, in practice, distinct methodological tools can be used to understand the same phenomenon in greater detail than any single paradigm could alone.

We begin this article by reviewing the history of reductionism and anti-reductionism. We then describe a third intermediate view, explanatory pluralism, which advocates the complementary use of more than one perspective at once, and has emerged as a way of studying complex systems in physics, biology, and other areas (Dale et al., 2012). This view, we argue, is especially well suited to the study of multiscale behavioral and cognitive phenomena (Ihlen and Vereijken, 2010; Kello et al., 2010; Dixon et al., 2012). We identify two benefits from practicing explanatory pluralism – *top-down constraining* and *bottom-up scaffolding* – and illustrate them through a case study of explanatory pluralism. We describe three empirical investigations of the same phenomenon at different levels of analysis, and from different theoretical perspectives. We end by considering how to critically assess and complement approaches, what is gained in this case by the pluralist approach, and what would be lost by more traditional reductive and non-reductive approaches.

MULTISCALE NATURE OF COGNITIVE AND BEHAVIORAL PHENOMENA

A clear example of the need for a plurality of approaches is to be found in the multiscale nature of cognitive and behavioral processes. Visual recognition happens through rapid millisecond dynamics of neural population codes in the brain (Mauk and Buonomano, 2004). However, precisely the way this happens is shaped on the longer timescales of ontogenesis and cultural evolution. For instance, sensitivity to certain color distinctions seems largely influenced by linguistic inheritance (Roberson et al., 2005; Winawer et al., 2007) and even the famous Müller-Lyer illusion has been found to be modulated by the saliency of carpentered corners in a given culture and environment: infants growing up in some cultures will be more prone to perceive all angles as square corners distorted by distance (Henrich and McElreath, 2003; Henrich et al., 2010). It is increasingly acknowledged that cognitive and behavioral phenomena generally involve multiple temporal and spatial scales (e.g., Newell, 1994; Dale et al., 2012).

As a working definition, we can define the scale of a method as the set of units typically used in analyses. On this definition the temporal scale of neural activity tends to emphasize a milliseconds-to-seconds range, while the temporal scale of geology tends to emphasize a kiloannums-to-gigannums (thousands to billions of years) range (cf. Newell's "Bands of Cognition": Newell, 1994). The spatial scale of a discipline or method relative to a phenomenon can be defined in a similar way. Neuroscience works mainly in the nanometer-to-centimeter scale, while ecology considers environments on a meter-to-kilometer scale. It has to be noted that a discipline or method can consider multiple spatial and or temporal scales, as well as relations between them: ecologists, for example, sometimes consider the relationship between relatively low-level chemical processes in the soil of a region and higher-level processes like the viability of species in that region; and of course, physics considers everything from the smallest scales of particle physics to the largest scales of the cosmos as a whole.

A prime example of a complex, multiscale cognitive and behavioral phenomenon is human language (Beckner et al., 2009). Units of language such as phonemes, syllables, words, phrases, texts, and discourse exist at distinct scales. They are studied at corresponding temporal and spatial scales, from raw acoustic energy patterns unfolding in the milliseconds range, to larger structures encompassing minutes, hours, and even days. The range extends further still, to the slower pace of language change and evolution that occurs over years and centuries. These different scales are studied using a variety of different frameworks and methods, including Fourier analysis, Markov chain analyses, discrete- and continuous-unit power law analyses, and for language in particular, corpus methods and semantic analyses. Linguistic behavior has been shown to be systematically organized across multiple time scales. Phonological distributions, word frequencies in a given language, and sequences of words in texts all follow power law distributions, where the frequencies of a given unit are in proportion across multiple scales of analysis (e.g., Zipf, 1949; Ferrer i Cancho et al., 2004; Kello and Beltz, 2009; Altmann et al., 2012).

As a consequence, we argue that no cognitive or behavioral phenomenon can be exhaustively described by reference to a single temporal or spatial scale or theoretical framework. The question thus should not be *which one scale* of analysis or *which one theoretical framework* is the right one to target and study for a given phenomenon. Rather, the issue is *which scales* and *which theoretical frameworks* are relevant for the question at hand, and *how they relate to each other*. This is the essence of the position called "explanatory pluralism." The alternative and more traditional account, which we delineate in more detail below, would be to focus on one scale of analysis and/or one theoretical framework for each given scale of phenomena.

EXPLANATORY PLURALISM

Contemporary philosophy of science grew out of the logical positivist movement of the 1920s, according to which the only meaningful statements are those that can be empirically verified. Psychology, for example, was taken to be meaningful only insofar as its statements could be translated in to the verifiable statements of a physical language (Carnap, 1959), and in that sense reduced to physics. This kind of view ran into various problems (e.g., the principle of verification seems to be meaningless according to its own criteria)[1], but the overarching project persisted: to understand in a formally rigorous way what science is and how different sciences are related to each other.

A standard view among the logical positivists, which remained even after positivism went out of fashion, was reductionism (Oppenheim and Putnam, 1958; Nagel, 1961; Cat, 2013 for review). According to the standard "layer cake" version of reductionism, higher-level "special sciences" (e.g., chemistry, biology) are arranged into a hierarchy, from physics to sociology, with physics at the bottom. It was thought that all the statements of any sciences besides physics could be reduced to the statements of the next lower level science via a system of "bridge laws." For example, one might assume that sociology would reduce to psychology, psychology would reduce to biology, biology would reduce to chemistry, and chemistry would reduce to physics. In this way, all empirical claims could ultimately be reduced to the laws of physics. This was the "reductionist" consensus until about 40 years ago[2]. It was also known as the "unity of science" view[3].

The more radical proponents of reductionism were also "eliminativists," who thought that, in light of ongoing reductions, all special sciences (psychology, economics, etc.) would be eliminated. In the end we would only need physics, because the other sciences are really just describing physical stuff using labels and other descriptive conveniences.

In the early 1970s the reductionist consensus came under attack. Fodor (1974), in an influential paper subtitled "the disunity of science," challenged reductionism by arguing that, even if it is true that nature is in some sense organized hierarchically, with fundamental particles aggregating into larger and larger systems of particles, this does not entail that higher level "special sciences" will be eliminated or reduced. The special sciences are, for Fodor, *not* eliminable; they are "autonomous." The reason is that higher-level theoretical vocabularies do not line up in tidy one-to-one ways with lower level theoretical vocabularies, the way the bridge-law approach suggested. A term like "desire" does not correspond one-to-one to a "natural kind" of neuroscience, since it is multiply realized in different kinds of organisms. Thus "desire" is a proprietary term of a distinctive science, psychology, which cannot be eliminated. Fodor's argument and other similar arguments (e.g., Davidson, 1969; see Cat, 2013 for review) were highly influential, and "non-reductive physicalism" became the new consensus for at least a decade.

A key feature of non-reductive-physicalism was the idea that, in practice, autonomous special sciences need not interact with lower-level sciences. Economists don't need to understand quantum field theory in order to study labor markets, even though labor markets are physical systems that obey the laws of fundamental physics. In fact, it would be a mistake, a waste of time, for an economist to consider such low-level phenomena. In a similar way, Fodor claims, psychology should describe laws of behavior

[1] The argument does not go through quite so easily, but there are still numerous problems with logical positivism and empiricism. For a detailed history and review, see Creath (2011) and Dienes (2008).

[2] Reductionism (as well as anti-reductionism and other positions we consider) can be understood as ontological or epistemological theses. As an ontological thesis, reductionism, for example, corresponds to the physicalist claim that there is nothing over and above physical entities. Our emphasis here is on the epistemological theses, which concern (for example) whether and to what extent knowledge and understanding at higher levels requires knowledge and understanding at lower levels.

[3] The concept of unity was actually understood in several, subtly different ways. See Cat (2013).

and cognition without wasting time on the low-level "implementation details" of neuroscience[4]. So non-reductive physicalism is associated with a kind of theoretical segregationism or "siloism" (our term), whereby different sciences are different levels, which maintain a principled isolation from one another.

So we have two views: (1) reductionism, where all special sciences reduce to physics, so that (in extreme eliminativist forms of this view) all sciences can in principle be eliminated except physics, and (2) anti-reductionism, where special sciences remain autonomous, and (in extreme "siloist" forms of this view) need not consult one another to do their work.

Explanatory pluralism is an intermediate third view, where special sciences are taken to be *semi*-autonomous (Edelman, 2008; Dale et al., 2009; de Jong, 2010; Hotton and Yoshimi, 2010; Yoshimi, 2012)[5]. On this view, different sciences have a degree of autonomy (they are not to be eliminated), but also interact in an effort to understand physical reality at different scales (they are not fully autonomous silos). According to the form of pluralism we advocate, different sciences and theoretical approaches should maintain their emphasis on different proprietary scales but should *also* work to unify their work as much as possible, insofar as they often describe the same phenomena in different but compatible ways.

Consider the old story about the blind men and an elephant (Saxe, 1884). Each of a group of blind men feels a different part of an elephant and then comes up with an incomplete, incompatible account of it.

> Six blind men encounter an elephant. Each feels a different part, and infers from the properties of the portion encountered the nature of the whole (one feels the tusk and concludes that he has encountered a spear, another feels the trunk and deduces that he has met a snake, etc.). It is often suggested that we are in the same position with respect to consciousness: different (even incompatible) theories may be derived from correct, but incomplete, views of reality. (Sloman and Chrisley, 2003, p. 4)

But with a little collaboration they can recognize that they are describing the same thing in different ways, and thereby collectively contribute to a fuller understanding of their target phenomenon.

Within cognitive science, variants of this pluralistic theme have a history, even if not by name. The concept of distinct "levels of analysis" goes back at least to Marr (1982), with his famous explanatory hierarchy of the computational, algorithmic, and implementational levels; each level with its own focused program of investigation. However, Marr's (1982) work was appropriated by Fodor and others to support strong forms of autonomy, which discourage interaction between theories at different levels. Just as software engineers don't need to understand the implementation details of the computer that runs the algorithms they write, so too psychologists don't need to understand the neural hardware that implements the algorithms and computations they describe[6].

Approaches advocating more pluralistic interactions between theories emerged in the 1980s and early 1990s, as researchers began to develop ways of unifying connectionist and symbolic approaches to cognition in common frameworks (e.g., Smolensky, 1988; Bechtel, 1990). More recently, a variety of theorists have developed frameworks for integrating different approaches to cognition. One example is the area of symbolic dynamics, where the lower-level dynamics of a system can be "coarse-grained" (multiple states at a lower level are treated as a single state at a higher level) and thereby analyzed in terms of discrete computational states (Dale and Spivey, 2005; Edelman, 2008; Yoshimi, 2010). These types of approaches allow researchers to study systems using multiple theoretical frameworks (e.g., dynamical systems theory and finite automata theory), but also to study the relationship between these theories (e.g., Crutchfield, 1994; Shalizi and Crutchfield, 2001; beim Graben and Potthast, 2009; Atmanspacher, 2011; Butterfield, 2011; Yoshimi, 2012; Dale and Vinson, 2013)[7].

Explanatory pluralism does not imply the anarchistic idea that "anything goes"[8]: often, more than one approach is needed, but not all approaches are equally motivated, and many are even not warranted. If two approaches contribute the same (or largely correlated) information about a phenomenon, they should be treated as competing alternatives. In such a case, either one will produce a better explanation (and the other is a mere symptom,

[4] Fodor's claim is that, assuming that a given pair of sciences (e.g., psychology and neuroscience) "cross-classify" the same phenomena, in the sense that they introduce predicates that do not map 1-to-1 onto each other (there is no isomorphism or set of "lawful coextensions" between them), then in practice it does not make sense for the two theories to interact (see Fodor, 1974, p. 113). Fodor clearly thinks it is a mistake to encourage cross-level interaction in the case of psychology and neuroscience: he bemoans the very idea of what would today be called "cognitive neuroscience": "There are departments of 'psycho-biology' and 'psychology and brain sciences' in universities throughout the world whose very existence is an institutionalized gamble that lawful co-extensions can be found" (Fodor, 1974, p. 105). Such attempts are "foredoomed." In another text Fodor and Pylyshyn argue at length against "brain-style modeling" in the cognitive sciences, and again treat it as a mistake, which they trace back to Lucretius: "the structure of 'higher levels' of a system are rarely isomorphic, or even similar, to the structure of 'lower levels' of a system. No one expects the theory of protons to look very much like the theory of rocks and rivers, even though, to be sure, it is protons and the like that rocks and rivers are 'implemented in'. Lucretius got into trouble precisely by assuming that there must be a simple correspondence between the structure of macrolevel and microlevel theories. . .it seems that the commitment to 'brain style' modeling leads to many of the characteristic Connectionist claims about psychology, and that it does so via the implicit and unwarranted-assumption that there ought to be similarity of structure among the different levels of organization of a computational system." (Fodor and Pylyshyn, 1988, p. 63). Though Fodor clearly abides by some kind of disciplinary isolation principle in such cases, subsequent non-reductive physicalists did not uniformly follow him in this (see, e.g., Sober, 1999).

[5] There are differences between these views, but our focus here is on a generic form of explanatory pluralism which captures the general idea that no one science is proprietary and that multiple sciences are needed to understand physical reality in all its complexity.

[6] We would not subscribe to this view of the relationship between software and hardware, at least in purist terms.

[7] Within pluralism, we find that explanatory pluralism and truth pluralism have potentially interesting relationships. For example, Lynch (2001) provides arguments for a metaphysical pluralism countering questions of absolutism. Considering our proposal for an explanatory pluralism that sits between reductionist and anti-reductionist views, there is room for ample discussion on how epistemic and ontological pluralism can fit together. We hold our views on the matter for a future paper, as we consider this important topic to be out of the scope of the current paper.

[8] Cf. Feyerabend (1975) ". . .there is only one principle that can be defended under *all* circumstances and in all stages of human development. It is the principle: anything goes."

which can be discarded), or it might turn out that they are both driven by a third factor that needs to be identified. A critical criterion for explanatory pluralism is thus that the multiple approaches should not only be motivated by complementary perspectives, but should also contribute different and independent (minimally correlating) information about the subject matter. The cumulative addition of approaches to a research question is only justified to the extent that each new approach enables the researcher to account for new aspects of the phenomenon that would be inaccessible given other approaches. Comparisons between approaches are also necessary in order to assess their reciprocal productivity and explanatory power. This can be done in at least three often related ways: (1) through a conceptual analysis of the approaches involved, (2) through a data-driven statistical model comparison, and (3) through a more direct experimental manipulation of the factors involved, aimed at disentangling the reciprocal role of the mechanisms suggested by the different models.

The current case study is an example of a conceptual analysis of explanatory pluralism. In this case there is no explicit model fitting or experimentation across levels, but rather a theoretical analysis of how multiple independently motivated analyses of the target phenomenon, framed at different temporal and spatial scales, are related to one another. If the role of the scientist is to investigate, observe, and continually add to explanations of phenomena, it seems obviously valuable to show how multiple observations and theories, despite differences in method and scale, can be complementary. A conceptual analysis can take different forms depending on the specific features (e.g., types of analysis) of the theories being integrated. The main idea is that there is a synthesis of results from various levels of analysis. The way to go about synthesizing depends in part on the type of analytic practice involved in the theories being synthesized, which we discuss later in this section.

A more direct way of applying explanatory pluralism is by using data-driven analysis. This requires the utilization of model-fitting procedures (e.g., stepwise linear regression indices such as adjusted R-squared, log-likelihood, AIC, BIC; see Schwarz, 1978; Hastie et al., 2009; see also Myung and Pitt, 1997) and, most importantly, commensurate units of data from multiple levels of analysis. Under this strategy, the question becomes: how much variance of the phenomenon does each level of analysis explain? Although this strategy might seem to be most optimal for "compatible" (Giere, 2004) levels of analysis, issues of measurement error and methodological assumptions can become limiting factors that need to be addressed. Related to data-driven practice is the experimental practice of carefully manipulating parameters in order to discriminate between "causal" roles of different mechanisms[9].

The third approach – designing experiments to test competing theories – is quite common in the realm of cognitive psychology, in which behavioral data can be leveraged against theoretical sticking points. By directly testing the predictions of potentially competing theories, an experimenter might confirm one theory or disconfirm another. This strategy is perhaps the most common approach to the theoretical sticking points in cognitive science. Famous recent examples include the "past-tense debate" (Pinker and Ullman, 2002) in psycholinguistics, or "prototypes vs. exemplars" in categorization research (Rouder and Ratcliff, 2006), in which dozens if not hundreds of empirical papers have explored these topics. In general, however, the degrees of freedom available to a theory, and to an experimenter, make it very difficult to develop "critical tests" and the weight of evidence on one side or the other has to gradually accumulate. Incidentally, neither of the vigorously pursued debates cited in this paragraph has been resolved to consensus, but integrative approaches have indeed been proposed for some (e.g., Love et al., 2004).

Explanatory pluralism affords the scientist a method for developing *fuller* explanations of relevant phenomena. The question then becomes how to apply explanatory pluralism in practice. In practice, what techniques are available for analyses and explanations of a phenomenon that exists at multiple temporal and spatial scales? Though there is no universally agreed upon model of explanation (Woodward, 2009), we can make a start by describing several specific approaches to explanatory pluralism: *top-down constraining* and *bottom-up scaffolding.*

Top-down constraining affords the scientist a basis for unifying multiple levels of analysis by identifying longer-scaled levels as *contextual constraints* for the smaller-scaled levels. For example, the amount of phonetic convergence (Pardo, 2006) – the phenomenon where the phonetic properties of interlocutors tend to align over the course of an interaction – depends on the contextual properties such as participant role and sex of the dyad. The contextual properties, such as the role of a participant in a conversational task, constrain behaviors occurring at shorter timescales such as the phonetic repertoire of interlocutors. We are not asserting a problematic "downward causality," but rather are describing a pattern of scientific practice. We are advocating that scientists identify and analyze the different types and levels of contextual influences on a phenomenon.

Bottom-up scaffolding provides a framework for identifying what can emerge from lower-level patterns (i.e., patterns existing at shorter time scales or smaller spatial scales), and the dynamics and processes by which these patterns are formed. It is the substrates of lower levels that allow higher-level phenomena to emerge. As with top-down constraining, we need not assert any kind of problematic cross-level causality. Bottom-up scaffolding provides the scientists with a means of expressing how (for example) symmetries at lower levels must be broken for distinct phenomena at higher levels to occur (Kugler and Shaw, 1990).

These ideas are inspired by the heuristic identity theory (HIT) proposed by McCauley and Bechtel (2001). In this theory, processes of bridging across levels of explanation are not a matter of simplistic isomorphism between laws, or mappings between ontologies. Instead, mapping across levels should create mutual constraint, in that levels should be consistent, if qualitatively, with each other. Mapping should also generate new questions, as each level may inspire new lines of investigation in the other. These two benefits of heuristic mapping may guide an eventual synergy between levels of analysis. "They enable scientists working at one analytical level to exploit the conceptual,

[9] Both these latter practices could be profitably applied to the case in analysis, but escape the scope of the current paper.

theoretical, and methodological and evidential resources available at another." (p. 743). HIT embraces both streams of influence proposed here: from top-down constraints and from bottom-up scaffolding.

Despite all this theoretical work supporting explanatory pluralism, there have been few if any detailed studies of specific cases. We fill this gap by considering a specific case in detail. In the case we consider, multiple frameworks are used to analyze the same data: a corpus of conversing individuals solving a joint decision-making task (Bahrami et al., 2010). We discuss three approaches: a systemic approach at the timescale of ~60–90 min in which the entire sequence of joint decisions is analyzed for its statistical properties, a lexical approach emphasizing the words spoken in the conversation at the timescale of minutes, and a physical approach focusing on the multiple time-scales of micro and macro coordination as expressed by the timescale of acoustic energy of participants' speech events. Each approach is born from very different theoretical assumptions, and focuses on a different scale using different theoretical and methodological tools. No single approach fully encompasses the phenomenon of joint decision-making. However, by taking all three approaches into account, we argue, joint decision-making is understood in a more articulated way than if it were studied at just one scale or using just one methodology. This is our notion of explanatory pluralism: the synergy of multiple theoretical frameworks targeting various scales of analysis in the investigation of a particular phenomenon[10].

CASE STUDY: JOINT DECISION-MAKING

Most of us must work in groups to complete complex tasks such as organizing conference symposia and collaborating on research projects; the production of this manuscript is one such example. In the past decade, a substantial research literature has emerged focusing on the cognitive, neurocognitive, behavioral, and physiological effects of working collectively in pairs or groups (for reviews see Fusaroli et al., in press; Pickering and Garrod, 2004; Shockley et al., 2009; Cooke et al., 2012). However, there is still much debate about whether individuals perform better than pairs or groups, and if so, how and under which conditions (e.g., Rajaram and Pereira-Pasarin, 2010).

Bahrami et al. (2010) recently developed a paradigm for studying collective perceptual decision-making that begins to address questions of joint perceptual performance. The paradigm was inspired by models of sensory integration that address how individuals integrate information from different sensory modalities (Ernst and Banks, 2002). Their goal was to test the question: Would two people be able to integrate their perceptual information, as individuals integrate information from different senses, in order to optimize their decisions? In other words, would two heads be better than one, and in particular, better than the best individual performance in a pair? They found that when two people were given the opportunity to communicate freely about their level of confidence on a trial-by-trial basis, two heads *became* better than one. However, this collaborative benefit was dependent on the interlocutors being equally good at solving the task on their own: differently performing interlocutors would not benefit from collaboration.

We argue that this joint decision-making paradigm provides a concrete case study for assessing explanatory pluralism. The three studies discussed are semi-autonomous in that they originate from disparate theoretical perspectives and focus on very different time scales, but at the same time complement each other, increasingly building an understanding of how and when interlocutors gain a collaborative benefit.

APPROACH 1: BEHAVIORAL/DECISION-MAKING (Bahrami et al., 2010)

In the original study (Bahrami et al., 2010), dyads were given a perceptual oddball task. The participants were recorded while sitting in front of their own respective screen at right angles to each other in a darkened room. The screens were identical and displayed exactly the same video output. On each trial the participants were sequentially shown two 85 ms long visual displays containing six Gabor patches. One of the displays would contain a contrast oddball: one of the six Gabor patches would have a stronger contrast and therefore look slightly darker (**Figure 1**).

The strength of the contrast varied randomly across trials. The participants were instructed to individually and separately indicate which of the displays contained this contrast oddball, by pressing a button. As long as both participants gave the same answer they would automatically proceed to the next individual trial. However, if their individual choices disagreed, they were prompted to negotiate, by freely discussing with each other, a joint decision. There

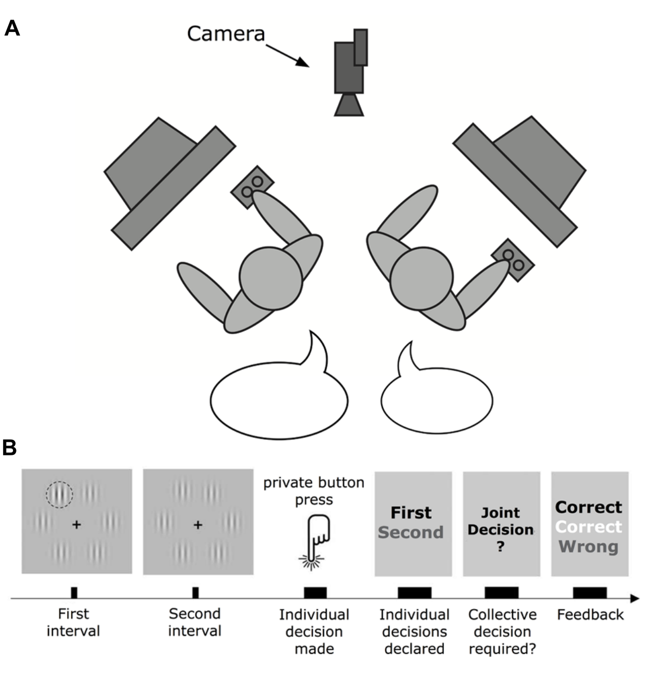


FIGURE 1 | Experimental setup (adopted with permission from Fusaroli et al., 2012). (A) The experimental setup. **(B)** Schematic illustration of a typical trial.

[10] Our review is inevitably selective. However, there are many variants of a pluralistic approach to science and cognition, in various domains, including: Abrahamsen and Bechtel (2006), Atmanspacher and beim Graben (2009), Dennett (1991), Dupré (1993), Eliasmith (1996), Kellert et al. (2006), Kelso and Enstrøm (2006), Mitchell (2003), Weiskopf (2009), among others.

was no time or other constraint on the joint decision dialogs. Individual and collective accuracy were then calculated by fitting a psychometric function to the dyad data[11]. The benefit of collaborating was then computed as the ratio between collective accuracy and the individual accuracy of the better of the two individuals.

Bahrami et al. (2010) used the empirical data thus produced to compare four models of information processing and transfer, each emphasizing different components of sensory processing, joint decisions, and communication: decision-making as relying on (1) a coin flip, (2) prioritizing the most perceptually competent group member's decision, (3) the sharing of confidence on the individual decisions, and (4) the sharing of the full perceptual information on the stimulus. The best explanation for the empirical data was model 3 – the weighted sharing of confidence on the individual decisions. However, the collaborative benefit was dependent on similarity of individual sensitivities to the stimuli contrasts: in other words, differently performing interlocutors would not benefit from collaboration.

From our perspective, the approach employed by Bahrami et al. (2010) required the coarse-grained aggregation of behaviors from every trial: The overall unit of analysis was the psychometric function calculated on the full sequence of joint decisions per each individual and per each pair. Finding that "two heads were better than one" required the aggregation of local behavioral responses (decision-making) into the global perceptual sensitivity – operationalized as the estimated psychometric functions for each individual and each dyad.

Aggregating over the dynamics of a given decision process is common to theoretical approaches in cognitive science. Indeed, testing the predictions of some theories *requires* such aggregation of outcomes, rather than of processes. Consider, for example, Bayesian accounts of cognition (Chater and Oaksford, 2003, 2008; Chater et al., 2006). Bayesian approaches consider *distributions* over potential decisions or states; the only way this can be achieved is by aggregating a large number of decisions or behaviors and characterizing their distribution. By doing so, we are able to use the Bayesian framework to predict the longer-term properties of a decision process, and assess whether that process obeys certain principles of rationality or optimality. In this way, the Bayesian approach and associated behavioral methods target specific levels of analysis, e.g., the purposive/computational of Marr's levels.

The thesis of the original Bahrami et al. (2010) paper has these same properties. In order to assess the overall optimality of a joint decision process, we must aggregate over perceptual decisions. The underlying dynamics of the decision process (which we consider below) seem less relevant here; we want to know whether participants were interacting, and whether the presence of interaction (as a discrete variable) shaped their joint accuracy in interesting ways. Put simply, these questions require us to point to certain aspects of a task, aggregate over these decisions, and assess the outcome of our analyses with respect to predictions from these frameworks. See **Figure 2** for a diagram of the Bahrami et al. (2010) paradigm, which shows how different levels of the task are studied by the different approaches considered here.

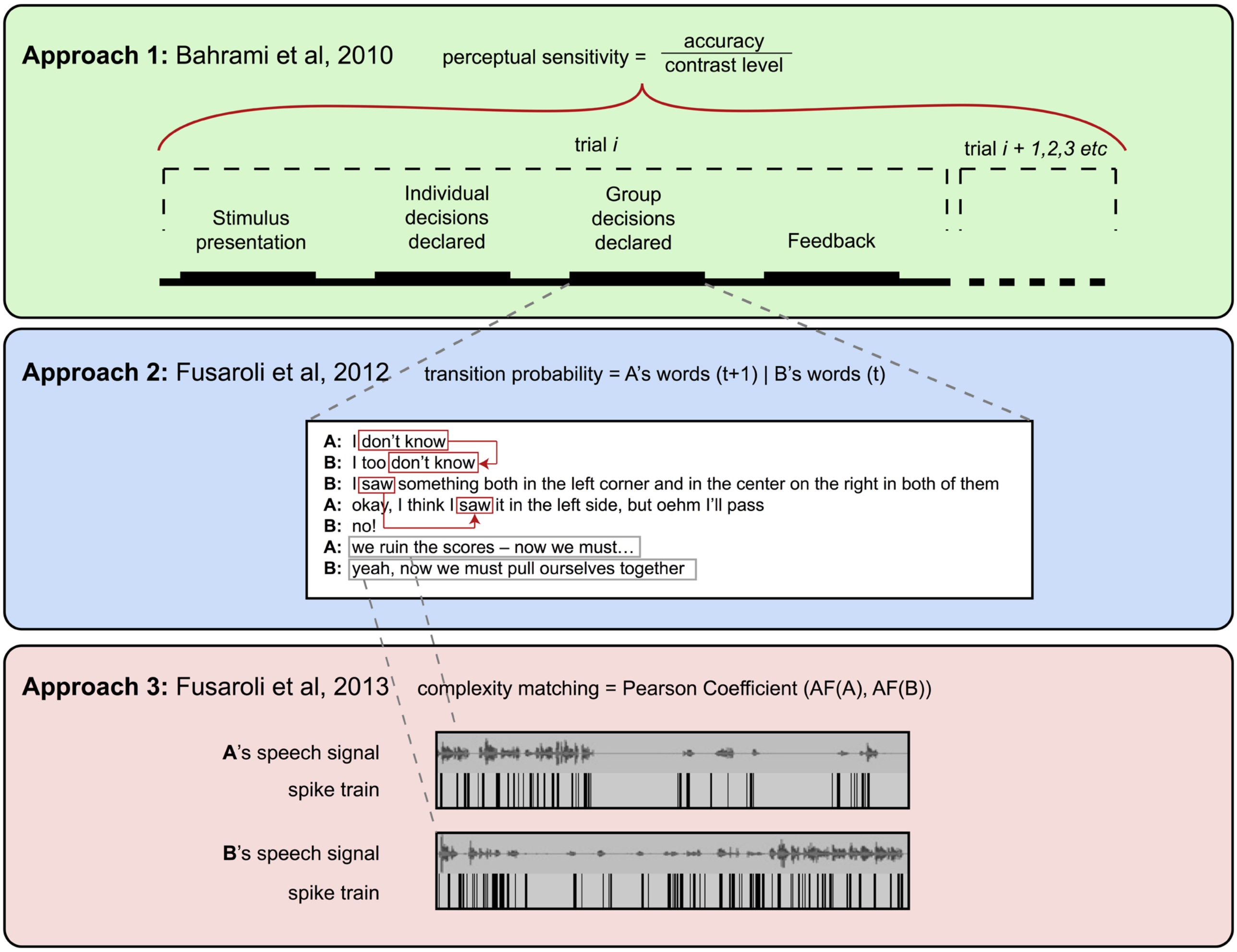
APPROACH 2: LINGUISTIC/CONFIDENCE (Fusaroli et al., 2012)

The second study we discuss had a very different starting point. During conversation, interlocutors have been observed to align to each other's linguistic behaviors (Pickering and Garrod, 2004; Fusaroli and Tylén, 2012). The degree of linguistic alignment has been shown to have functional value; for example, high linguistic alignment tends to assist in some contexts of problem solving (Garrod and Anderson, 1987; Garrod and Doherty, 1994). However, it is disputable whether linguistic alignment is always beneficial and in what ways (for instance the extreme case of echolalia, where one interlocutor simply repeats what the other says, does not seem to be an effective conversational strategy; Fusaroli et al., 2014). Fusaroli et al. (2012) re-analyzed the Bahrami et al. (2010) experiment to explicitly investigate differences in conversational strategies employed by well- and poorly performing dyads. The aim was to map which aspects of linguistic alignment were functional for group performance in the joint decisions. This perspective required a finer-grained look at the actual process of decision-making, and not just on its results.

The videos of the joint decision-making tasks were transcribed. Following on Bahrami et al.'s (2010) insights that confidence sharing is crucial to effectively solve the task, the transcripts were coded for participants' spontaneous ways of sharing confidence linguistically, for instance through expressions such as "I think I saw something." Across the 16 dyads analyzed, 35 types of such confidence expressions were identified, e.g., modulations of "to see" as opposed to "to be sure." Distributional patterns and token frequencies for each type of confidence expression were quantified in each transcript. *Local linguistic alignment* was calculated for all lexical items as the transitional probability of any given expression used by one participant being used by the other participant in the preceding joint decision. For example, local linguistic alignment was computed as the probability of Participant B using the expression "to see" when Participant A used the same expression in the previous trial.

By having indices of different types of linguistic alignment – one type that subsumes all lexical expressions (indiscriminate alignment) and one that subsumes only confidence expressions (discriminant alignment) – Fusaroli et al. (2012) were able to determine which type of alignment benefitted joint decision-making: task-specific, discriminate alignment or general, indiscriminate alignment. To measure the lasting effects of alignment, Fusaroli et al. (2012) also calculated a more coarse-grained measure of *global linguistic convergence* of confidence expressions. This was computed as the percentage of the overall sum of confidence tokens used by the dyad throughout the experiment, which belonged to the most frequent confidence type thus abstracting from the local linguistic exchanges between interlocutors and focusing on long term linguistic consistency of the pair.

[11] Psychometric functions were computed by plotting the proportion of trials in which the oddball match was reported in the second interval, as a function of the contrast difference with the surrounding patch array. The functions were fit with a cumulative Gaussian function. The slope of each function provided an estimate of perceptual sensitivity: the steeper the slope, the higher the sensitivity.


Approach 1: Bahrami et al, 2010
perceptual sensitivity = $\frac{\text{accuracy}}{\text{contrast level}}$
trial *i*
trial *i + 1,2,3 etc*
Stimulus presentation
Individual decisions declared
Group decisions declared
Feedback
Approach 2: Fusaroli et al, 2012
transition probability = A's words (t+1) | B's words (t)
Approach 3: Fusaroli et al, 2013
complexity matching = Pearson Coefficient (AF(A), AF(B))
A's speech signal
spike train
B's speech signal
spike train

A: I don't know
B: I too don't know
B: I saw something both in the left corner and in the center on the right in both of them
A: okay, I think I saw it in the left side, but oehm I'll pass
B: no!
A: we ruin the scores – now we must...
B: yeah, now we must pull ourselves together

FIGURE 2 | The three approaches to decision-making paradigm discussed here, with an indication of their characteristic temporal and spatial scales as well as their favored methodological tools.

Relying on Bahrami et al.'s (2010) psychometric function for calculating collective performance, Fusaroli et al. (2012) observed that task-specific, local linguistic alignment and global linguistic convergence positively correlated with collective benefit, whereas, local indiscriminate alignment negatively correlated with collected benefit. Furthermore, global convergence strongly predicted collective benefit: when dyads continually and consistently used shared sets of linguistic expressions for expressing confidence, collective benefit was observed to be higher. Fusaroli et al. (2012) concluded that in order for dyad members to benefit from cooperation they should not just parrot each other (indiscriminate alignment). Rather dyads that jointly adapted linguistic tools to meet the functional affordances of the task (sharing and comparing confidence) reached high collective performance.

Summing up, the search for functional linguistic alignment led Fusaroli et al. (2012) to conduct a corpus analysis of trial-to-trial transcripts highlighting the actual communication strategies employed to solve the joint decision task. Computing the transitional probabilities of lexical alignment required a fine-grained analysis of the local dynamics of lexical choices – the unit of analysis being lexical expressions within adjacent joint decisions – keeping track of individuals' productions. It has to be noted that coarse-grained analyses were also crucial for the study: global linguistic convergence, collective benefit and even the aggregate measures of linguistic alignment to be used as dyad-level correlates for collective benefit. Importantly, the aggregation procedure applies at a different level of the analysis – at the finer-grained timescale of words being used by interlocutors.

As described above, aggregation in Bahrami et al. (2010) derived from a desire to quantify coarser-grained characteristics of the decision process: psychophysical sensitivity and the benefits of interaction. The Fusaroli et al. (2012) study provided insight into the mechanism and process of collective benefit, whereas the Bahrami et al. (2010) study provided a description of *why* these mechanisms work in a particular way. But, as discussed in background sections above, every theory has boundary conditions that limit the claims it can make about complex systems. Put simply, the aggregation approach is unable to specify both *how and why* the content or structure of an interaction helps joint decisions. Trying to get at these new aspects invokes more dynamic linguistic and psycholinguistic theoretical machinery, quickly leading to new questions and *different* levels of analysis. Instead of aggregating decisions alone, we "peel back" those decisions and peer into their contents, before aggregating in different ways from before.

APPROACH 3: PHYSICAL/ACOUSTIC ENERGY (Fusaroli et al., 2013)

Recent work on "complexity matching" in the field of statistical physics has shown that information transmission between two complex systems is optimal when the complexities of the behaviors of the two systems match (West et al., 2008). Fusaroli et al. (2013) investigated whether such ideas would apply to human interactions (see also Abney et al., under review). Indeed, it has been shown that humans produce behaviors with long-range correlations at increasing time scales, and additionally, behaviors are observed to follow scaling laws evidenced by heavy-tailed distributions (Kello et al., 2010). Fusaroli et al. (2013) thus investigated if the statistical complexities of behaviors would match between two interacting humans, and if so, according to the physical models, the degree of complexity matching would predict information transfer to be optimal. Already Fusaroli et al. (2012) were looking for a functional relationship between the degree of matching of a particular behavior and the accompanying performance outcome; however, the use of complex systems physical models creates important divergences in the level of description and in the time scales of subsequent analyses.

To estimate the multiscale complexity matching of human behavior during interactions, Fusaroli et al. (2013) had to employ yet a new unit of analysis, capturing more basic perceptuomotor coupling between interlocutors. For this new unit they had to first assess the complexity (hierarchical scaling; cf. Abney et al., under review) and then the match in complexity between interlocutors, with the hypothesis that the more the complexity of participants' speech behavior matched, the higher the collaborative benefit. They analyzed the physical basis or "skeleton" of linguistic exchanges: the acoustic energy of speech events of individuals in conversation. Onset/offset intervals of the acoustic signal from the conversations were extracted by identifying boundaries between speech and pauses (pauses were defined as reduced acoustic intensity and the absence of pitch lasting beyond 20 ms). Binary spike trains of speech events were computed from the onset/offset intervals; states were coded with "0s"; "1s" were used to code changes in state, that is, the onset or offset of a speech event. Thus, the unit of analysis was the onset/offset of a speech event defined by the presence or absence of particular properties of acoustic energy. A temporal estimate of complexity of human behavior was computed for each participant and each joint conversation trial employing Allan Factor (AF, Allan, 1966), a multiscale method for estimating the correlated clustering of speech events[12]. Complexity matching was defined as the degree of correlation between AF estimates of participants in a dyad. Across all trials, the authors found a positive correlation between degree of complexity matching and collective benefit: when the complexities of participants' speech behaviors matched, collective benefit on the joint perceptual decision task was higher. These findings can be interpreted as preliminary evidence for complexity matching in interpersonal coordination (West et al., 2008): Increased collective benefit in a dyad can be considered an index of the optimality of information transfer, which increased as a function of the coordination of human behavior across multiple time scales. In other words, the more fine-tuned the turn-taking coordination of the interlocutors, the better the information transfer, which in turn led to a higher collective benefit. Crucially, the degree of complexity matching increased from the first to the second half of the experiment, suggesting that complexity matching express the degree to which interlocutors adapt to coordinate with each other.

This third study used a trial-by-trial measure that summarized the multiscale properties of speech coordination in its very basic form of acoustic energy. The overall unit of analysis was the onset/offset of acoustic energy at a 10 ms time scale. This fine-grained analysis was then aggregated into a coarse-grained analysis that operationalized the degree of complexity matching between dyads. The degree of complexity matching was then correlated with collective benefit computed for either all trials or session-by-session.

Again, the questions regarding the microstructure of coordination cannot come from linguistic analysis or from aggregation of perceptual decisions. Instead, it starts from a much more fine-grained level of analysis, specifically the dynamics of the perceptuomotor structure of the task. Just as theories about optimal decisions or linguistic alignment require selecting particular levels of description and analysis, here researchers choose a finer timescale and extract signals, which may be subject to their own (but different) aggregation. Researchers who adopt this theory often require very densely sampled behaviors to quantify and characterize the dynamics that underpin some behavior or cognitive process. Now quite different observations can be made about the process of interaction, regarding the composition of the task in terms of the perceptuomotor coupling of its participants.

The past three sections laid out sample re-analyses of the Bahrami et al. (2010) dataset, with quite different goals in mind. The upshot of this research scenario is to select disparate scales of analysis, different means of measurement, and different patterns of aggregation in order to assess particular predictions about the cognitive system, or to characterize the cognitive system in different ways at different levels. However, we do not mean to suggest that these are completely independent levels of analysis (*à la* Fodor, 1974, 1975; Fodor and Pylyshyn, 1988). Instead, these processes should be seen as interdependent, inspiring each other and providing reciprocal insight. Indeed, Fusaroli et al. (2012, 2013) already shows hints of this integration: The optimality of the joint decision process can be correlated with the structure of the linguistic interaction; similarly, the joint decision process and linguistic structure itself may be related to the dynamics of the acoustic speech behaviors of dyad members, such as the correlation

[12] The Allan factor analysis computes the correlation estimate – α – of the variance of speech events at a particular time scale across multiple time scales. A scaling relation, or power law, of speech events is evidenced when $\alpha \sim 1$ whereas, and $\alpha \sim 0$ is considered a Poisson process. The complexity of a participant's speech behavior is determined by the computation of α.

between complexity matching indices. In the next section, we discuss this potential integration more fully, noting that explanatory pluralism also encourages this kind of synthesis across theoretical domains.

SYNTHESIS OF LEVELS OF DESCRIPTION

Explanatory pluralism has been presented as a view intermediate between extreme forms of reductionism (where everything ends up being physics) and anti-reductionism or strong autonomy (where different sciences are insulated from one another). Above, we have seen a detailed example of explanatory pluralism in practice, with three different studies approaching the same phenomenon – joint decision-making by a pair of participants engaged in a perceptual discrimination task – at different temporal and spatial scales, using distinct methodological tools.

In this section, we return to the topic of explanatory pluralism, and consider how this kind of approach works in practice. First we discuss the benefits of each level of description. Then we discuss advantages of interactions between these levels, and how syntheses between them can motivate new questions and insights.

BENEFITS OF INDEPENDENT LEVELS OF DESCRIPTION

Against views which emphasize the value of a single paradigm in cognitive science (e.g., reductionism in its most radical form, which advocates focusing only on the physical level), explanatory pluralism holds that it is important to study phenomena using multiple independent levels of analysis. We have seen how the three levels discussed above are important to understanding performance in joint decision-making. Bahrami et al. (2010) found that, when perceptual sensitivities were equal, dyads benefitted from interaction by comparing levels of confidence. Fusaroli et al. (2012) observed that group performance was higher when interlocutors shared common task-relevant lexical patterns during conversations about confidence. Finally, Fusaroli et al. (2013) provided evidence that group performance increased when the hierarchical structure of the very basic patterns of interlocutors' vocalizations – with the base unit of the onset of acoustic energy – matched within the dyad. These insights were obtained while remaining within the relevant spatial and temporal scale and while making use of the characteristic tools of each approach.

What if a more traditional approach were followed, which only allowed for one way of analyzing joint decision-making? What would be lost?

If "Approach 1" were pursued in isolation, we would not know that the development and sharing of a linguistic confidence scale among members of a dyad makes a difference in the performance in the task (Fusaroli et al., 2012). Furthermore, the performance of the task is also successfully predicted by the hierarchical structure of acoustic energy onsets (Fusaroli et al., 2013).

If "Approach 2" were pursued in isolation, we would not know that the degree of matching in individual sensitivity is an important drive in the efficacy of confidence sharing. Additionally, we would not consider that the patterns of matching found in local linguistic alignment might be complemented by more basic patterns of matching of the hierarchical structures of acoustic energy onsets.

Finally, if "Approach 3" were pursued in isolation, we would not know that the rate of indiscriminate matching of lexical items negatively predicts performance; the alignment of language not functionally relevant for the particular context, does not help the dyad. Additionally, we would never be able to consider the possibility that unequal perceptual capabilities of individuals might have a significant effect on group tasks and relatedly, how, in turn, these asymmetries might affect linguistic-level and acoustic-level matching dynamics.

These considerations support the idea that strong forms of reductionism, which at times suggest outright elimination of all but the lowest level physical sciences, are problematic. All three approaches shed light on important features of joint decision-making. Eliminating any of these approaches leaves important features of the phenomenon unexplained. The synthesis of these three levels provides a more complete description of how people work together to solve a particular task.

BENEFITS OF INTERDEPENDENT LEVELS OF DESCRIPTION

Against views like strong *anti-reductionism* or "siloism", which advocate that different levels of analysis be *completely* autonomous, we believe that multiple theories should interact when describing the same phenomenon (cf. Simon, 1992). To make a case for this idea, we first consider what would be missing from a description of joint perceptual decision-making if the different approaches did not interact with each other.

If none of these three approaches informed each other, various research opportunities crossing scales and mixing methods would be lost. For example, can people with asymmetric perceptual capabilities effectively overcome their difference by communicating about common environmental constraints (Approach 1) and what do the language properties (Approach 2) look like when they successfully coordinate? Additionally, does local linguistic alignment or global convergence of linguistic coordination (Approach 2) relate to the matching of hierarchical structures at the level of acoustic energy onsets (Approach 3)? Finally, how does the degree of matching of hierarchical structures of acoustic energy onsets (Approach 3) relate to different dyad-level perceptual asymmetries (Approach 1)? All of these questions pertain to the corpus of data described in the present case study, and indeed some of these questions are currently being tested. What is important to understand is that all three levels are describing the same phenomenon, and there are certainly more levels of description that can be included. Asking these cross-level questions might persuade some to argue for reductionism, e.g., "does linguistic alignment just merely reduce to complexity matching?" or "how does linguistic alignment interact with dyad-level perceptual asymmetries?" We suggest that there is *interdependence* across levels, where theories can inform each other, ultimately leading to a better understanding of the phenomenon. Having an epistemological process prioritizing the interdependence across different levels stands in contrast to views suggesting that each level is independent, autonomous and represents competing explanations of a phenomenon and that higher levels can be reduced to lower levels.

TOPICS FOR FURTHER WORK IN EXPLANATORY PLURALISM

We have argued for explanatory pluralism using a detailed case study. However, it is important to point out that not all levels of description are complementary, and the principle of "more levels of description is better" is problematic if applied haphazardly and without a proper supporting framework. While a plurality of approaches is necessary to better explain a given phenomenon, not all approaches are equal: At what point does the diffusion toward a troubled eclecticism stop? We have earlier suggested three methods: (1) conceptual analysis, (2) data-driven statistical model comparison, and (3) experimental manipulations. This paper is an example of the first: we have assessed three approaches to the same phenomenon (and dataset) articulating their differences and complementarity. A data-driven statistical model comparison – which is outside of the scope of this paper – would comparatively assess the relation between the models and between the models and the data. In other words, it would look at the comparative explanatory power of individual performance, linguistic alignment and complexity matching: Do they equally fit the data? Do they explain comparable amount of variance in the data? Do we get better statistical fit and explanatory power by integrating individual performance, linguistic alignment and complexity matching in the same model and how do they interact? An experimental approach could push these comparisons further by investigating the causal relations between parameters and opening new venues of inquiry. For example, if we systematically vary one of the parameters (e.g., similarity in individual performance, by introducing noise in the stimuli presented to the worse performer in a dyad), how do the others vary in their levels and relation to performance?

In conclusion, if the goal of a scientific endeavor (e.g., understanding decision-making of pairs of humans) is to fully understand and thus predict future behaviors, then taking into account multiple levels and theoretical approaches is warranted if not necessary.

We have provided a conceptual treatment of what explanatory pluralism looks like in cognitive science. Although beyond the scope of the current paper, we also advocate the use of data-driven statistical model comparison and experimental manipulations to critically assess the interest and complementarity of different approaches. Indeed, some of us (Fusaroli and Tylén, under review) have already implemented this technique.

Additionally, we introduced two benefits of *practicing* explanatory pluralism in scientific investigations: *top-down constraining* and *bottom-up scaffolding*. Both benefits provide frameworks that may lead to future research questions about a particular phenomenon. Again, top-down constraining unifies multiple levels of analysis by identifying the longer-scaled levels as contextual constraints for the smaller-scaled levels. For example, Approach 1 provided contextual constraints for the linguistic tools (Approach 2) participants might utilize, and therefore, the patterns of acoustic energy (Approach 3). This framework affords the identification of contextual influences of a phenomenon not otherwise integrated across multiple levels of analysis.

Bottom-up scaffolding can be used to identify what can emerge from lower-level patterns. For example, the lexical items participants jointly use and align (Approach 2) emerge from the multiscale patterns of acoustic energy (Approach 3). Furthermore, an optimal model of joint perceptual decision-making, requiring the sharing of confidence (Approach 1), is comprised of the lexical items (Approach 2). Again, it is the substrates of lower levels of analysis that afford the possibility for higher-level components – via higher levels of analysis – of a phenomenon to emerge.

CONCLUSION

We have defended explanatory pluralism using a case study, which involved three separate analyses of the same phenomenon. We have made the case that integrating data and theory from multiple scales of analysis provides a fuller explanation of a cognitive phenomena than would be possible if we pursued a more traditional, theoretically autonomous style of scientific investigation.

Our call is for more researchers in the cognitive and behavioral sciences to consider studying phenomena of interest using the framework of explanatory pluralism. This can encompass a variety of practices ranging from conceptual analysis to full fledged, data-driven analysis. Acknowledging that theoretical approaches influence methodological decisions and practices, we argue that explanatory pluralism might be beneficial to the ultimate scientific endeavor of explanation.

REFERENCES

Abrahamsen, A., and Bechtel, W. (2006). "Phenomena and mechanisms: putting the symbolic, connectionist, and dynamical systems debate in broader perspective," in *Contemporary Debates in Cognitive Science*, ed. R. Stainton (Oxford: Basil Blackwell).

Allan, D. W. (1966). Statistics of atomic frequency standards. *Proc. IEEE* 54, 221–230. doi: 10.1109/PROC.1966.4634

Altmann, E. G., Cristadoro, G., and Degli Esposti, M. (2012). On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11582–11587. doi: 10.1073/pnas.1117723109

Atmanspacher, H. (2011). "Quantum approaches to consciousness," in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Stanford, CA: Stanford University).

Atmanspacher, H., and beim Graben, P. (2009). Contextual emergence. *Scholarpedia* 4, 7997. doi: 10.4249/scholarpedia.7997

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. (2010). Optimally interacting minds. *Science* 329, 1081–1085. doi: 10.1126/science.1185718

Bechtel, W. (1990). Multiple levels of inquiry in cognitive science. *Psychol. Res.* 52, 271–281. doi: 10.1007/BF00877535

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., et al. (2009). Language is a complex adaptive system: position paper. *Lang. Learn.* 59, 1–26. doi: 10.1111/j.1467-9922.2009.00533.x

beim Graben, P., and Potthast, R. (2009). Inverse problems in dynamic cognitive modeling. *Chaos* 19, 015103. doi: 10.1063/1.3097067

Butterfield, J. (2011). Less is different: emergence and reduction reconciled. *Found. Phys.* 41, 1065–1135. doi: 10.1007/s10701-010-9516-1

Carnap, R. (1959). "The Elimination of metaphysics through logical analysis of language," in *Logical Positivism*, ed. A. J. Ayer (New York: Macmillan), 60–81.

Cat, J. (2013). "The unity of science," in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta. Available at: http://plato.stanford.edu/archives/sum2013/entries/scientific-unity/

Chater, N., and Oaksford, M. (2003). *Rational Models of Cognition. Encyclopedia of Cognitive Science.* Hoboken, NJ: John Wiley and Sons.

Chater, N., and Oaksford, M. (eds). (2008). *The Probabilistic Mind: Prospects for Bayesian Cognitive Science.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199216093.001.0001

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: conceptual foundations. *Trends Cogn. Sci.* 10, 287–291. doi: 10.1016/j.tics.2006.05.007

Cooke, N. J., Gorman, J. C., Myers, C. W., and Duran, J. L. (2012). Interactive team cognition. *Cogn. Sci.* 37, 255–285. doi: 10.1111/cogs.12009

Creath, R. (2011). "Logical Empiricism," in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Stanford, CA: Stanford University).

Crutchfield, J. P. (1994). The calculi of emergence: computation, dynamics, and induction. *Physica D* 75, 11–54. doi: 10.1016/0167-2789(94)90273-9

Dale, R., Dietrich, E., and Chemero, A. (2009). Explanatory pluralism in cognitive science. *Cogn. Sci.* 33, 739–742. doi: 10.1111/j.1551-6709.2009.01042.x

Dale, R., and Spivey, M. J. (2005). From apples and oranges to symbolic dynamics: a framework for conciliating notions of cognitive representation. *J. Exp. Theor. Artif. Intell.* 17, 317–342. doi: 10.1080/09528130500283766

Dale, R., Tollefsen, D. P., and Kello, C. T. (2012). "An integrative pluralistic approach to phenomenal consciousness," in *Being in Time: Dynamical Models of Phenomenal Experience,* eds S. Edelman, T. Fekete, and N. Zach (Amsterdam: John Benjamins), 231–258.

Dale, R., and Vinson, D. W. (2013). The observer's observer's paradox. *J. Exp. Theor. Artif. Intell.* 25, 303–322. doi: 10.1080/0952813X.2013.782987

Davidson, D. (1969). "The individuation of events," in *Essays in Honor of Carl G. Hempel*, ed. N. Rescher (Dordrecht: Riedel), 216–234. doi: 10.1007/978-94-017-1466-2_11

de Jong, H. L. (2010). From theory construction to deconstruction the many modalities of theorizing in psychology. *Theory Psychol.* 20, 745–763. doi: 10.1177/0959354310376428

Dennett, D. C. (1991). Real patterns. *J. Philos.* 88, 27–51. doi: 10.2307/2027085

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference.* New York: Palgrave Macmillan.

Dixon, J. A., Holden, J. G., Mirman, D., and Stephen, D. G. (2012). Multifractal dynamics in the emergence of cognitive structure. *Top. Cogn. Sci.* 4, 51–62. doi: 10.1111/j.1756-8765.2011.01162.x

Dupré, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science.* Cambridge: Harvard University Press.

Edelman, S. (2008). On the nature of minds, or: truth and consequences. *J. Exp. Theor. Artif. Intell.* 20, 181–196. doi: 10.1080/09528130802319086

Eliasmith, C. (1996). The third contender: a critical examination of the dynamicist theory of cognition. *Philos. Psychol.* 9, 441–463. doi: 10.1080/09515089608573194

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi: 10.1038/415429a

Ferrer i Cancho, R., Solé, R. V., and Köhler, R. (2004). Patterns in syntactic dependency networks. *Phys. Rev. E* 69, 051915. doi: 10.1103/PhysRevE.69.051915

Feyerabend, P. K. (1975). *Against Method.* London: New Left Books.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28, 97–115. doi: 10.1007/BF00485230

Fodor, J. A. (1975). *The Language of Thought.* Cambridge, MA: Harvard University Press.

Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71. doi: 10.1016/0010-0277(88)90031-5

Fusaroli, R., Abney, D. H., Bahrami, B., Kello, C. T., and Tylén, K. (2013). Conversation, coupling, and complexity: matching scaling laws predicts performance in a joint decision task. *Poster Presented at the 35th Annual Conference of the Cognitive Science Society*. Berlin.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., et al. (2012). Coming to terms quantifying the benefits of linguistic coordination. *Psychol. Sci.* 23, 931–939. doi: 10.1177/0956797612436816

Fusaroli, R., Konvalinka, I., and Wallot, S. (in press). "Analyzing social interactions: potentialities, and challenges of CRQA," in *Springer Proceedings for Mathematics.*

Fusaroli, R., Rączaszek-Leonardi, J., and Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas Psychol.* 32, 147–157. doi: 10.1016/j.newideapsych.2013.03.005

Fusaroli, R., and Tylén, K. (2012). Carving language for social coordination: a dynamical approach. *Interact. Stud.* 13, 103–124. doi: 10.1075/is.13.1.07fus

Garrod, S., and Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition* 27, 181–218. doi: 10.1016/0010-0277(87)90018-7

Garrod, S., and Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition* 53, 181–215. doi: 10.1016/0010-0277(94)90048-5

Giere, R. N. (2004). How models are used to represent reality. *Philos. Sci.* 71, 742–752. doi: 10.1086/425063

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Linear Methods for Regression.* New York: Springer.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83. doi: 10.1017/S0140525X0999152X

Henrich, J., and McElreath, R. (2003). The evolution of cultural evolution. *Evol. Anthropol.* 12, 123–135. doi: 10.1002/evan.10110

Hotton, S., and Yoshimi, J. (2010). The dynamics of embodied cognition. *Int. J. Bifurcat. Chaos* 20, 943–972. doi: 10.1142/S0218127410026241

Ihlen, E. A., and Vereijken, B. (2010). Interaction-dominant dynamics in human cognition: beyond $1/f^{\alpha}$ fluctuation. *J. Exp. Psychol. Gen.* 139, 436. doi: 10.1037/a0019098

Kellert, S. H., Longino, H. E., and Waters, C. K. (2006). "Introduction: the Pluralist Stance," in *Scientific Pluralism, Studies in the Philosophy of Science*, Vol. 19, eds S. H. Kellert, H. E. Longino, and C. K. Waters (Minneapolis, MN: University of Minnesota Press), 7–29.

Kello, C. T., and Beltz, B. C. (2009). "Scale-free networks in phonological and orthographic wordform lexicons," in *Approaches to Phonological Complexity*, eds I. Chitoran, C. Coupé, E. Marsico, and F. Pellegrino (Berlin: Mouton de Gruyter).

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., et al. (2010). Scaling laws in cognitive sciences. *Trends Cogn. Sci.* 14, 223–232. doi: 10.1016/j.tics.2010.02.005

Kelso, J. A. S., and Enstrøm, D. (2006). *The Complementary Nature.* Cambridge: MIT Press.

Kugler, P. N., and Shaw, R. E. (1990). "Symmetry and symmetry-breaking in thermodynamic and epistemic engines: a coupling of first and second laws," in *Synergetics of Cognition*, ed. H. Haken (Heidelberg: Springer), 296–331.

Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychol. Rev.* 111, 309. doi: 10.1037/0033-295X.111.2.309

Lynch, M. P. (2001). *Truth in Context: An Essay on Pluralism and Objectivity.* Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Mauk, M. D., and Buonomano, D. V. (2004). The neural basis of temporal processing. *Annu. Rev. Neurosci.* 27, 307–340. doi: 10.1146/annurev.neuro.27.070203.144247

Mitchell, S. (2003). *Biological Complexity and Integrative Pluralism.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511802683

Nagel, E. (1961). *The Structure of Science.* London: Routledge and Kegan Paul.

Newell, A. (1994). *Unified Theories of Cognition*, Vol. 187. Cambridge, MA: Harvard University Press.

McCauley, R. N., and Bechtel, W. (2001). Explanatory pluralism and heuristic identity theory. *Theory Psychol.* 11, 736–760. doi: 10.1177/0959354301116002

Myung, I. J., and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon. Bull. Rev.* 4, 79–95. doi: 10.3758/BF03210778

Oppenheim, P., and Putnam, H. (1958). "Unity of science as a working hypothesis," in *Minnesota Studies in the Philosophy of Science*, Vol. 2, eds H. Feigl, G. Maxwell, and M. Scriven (Minneapolis, MN: University of Minnesota Press), 3–36.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.* 119, 2382–2393. doi: 10.1121/1.2178720

Pickering, M. J., and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–189. doi: 10.1017/S0140525X04000056

Pinker, S., and Ullman, M. T. (2002). The past and future of the past tense. *Trends Cogn. Sci.* 6, 456–463. doi: 10.1016/S1364-6613(02)01990-3

Port, R. F., and Van Gelder, T. (eds). (1995). *Mind as Motion: Explorations in the Dynamics of Cognition.* Cambridge, MA: MIT Press.

Rajaram, S., and Pereira-Pasarin, L. P. (2010). Collaborative memory: cognitive research and theory. *Perspect. Psychol. Sci.* 5, 649–663. doi: 10.1177/1745691610388763

Roberson, D., Davidoff, J., Davies, I. R., and Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. *Cogn. Psychol.* 50, 378–411. doi: 10.1016/j.cogpsych.2004.10.001

Rouder, J. N., and Ratcliff, R. (2006). Comparing exemplar-and rule-based theories of categorization. *Curr. Dir. Psychol. Sci.* 15, 9–13. doi: 10.1111/j.0963-7214.2006.00397.x

Saxe, J. G. (1884). *The Poems of John Godfrey Saxe.* Houghton: Mifflin and Company.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Shalizi, C. R., and Crutchfield, J. P. (2001). Computational mechanics: pattern and prediction, structure and simplicity. *J. Stat. Phys.* 104, 817–879. doi: 10.1023/A:1010388907793

Shockley, K., Richardson, D. C., and Dale, R. (2009). Conversation and coordinative structures. *Top. Cogn. Sci.* 1, 305–319. doi: 10.1111/j.1756-8765.2009.01021.x

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychol. Sci.* 3, 150–161. doi: 10.1111/j.1467-9280.1992.tb00017.x

Sloman, A., and Chrisley, R. (2003). Virtual machines and consciousness. *J. Conscious. Stud.* 10, 4–5.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behav. Brain Sci.* 11, 1–23. doi: 10.1017/S0140525X00052432

Sober, E. (1999). The multiple realizability argument against reductionism. *Philos. Sci.* 542–564. doi: 10.1086/392754

Weiskopf, D. A. (2009). The plurality of concepts. *Synthese* 169, 145–173. doi: 10.1007/s11229-008-9340-8

West, B. J., Geneston, E. L., and Grigolini, P. (2008). Maximizing information exchange between complex networks. *Phys. Rep.* 468, 1–99. doi: 10.1016/j.physrep.2008.06.003

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7780–7785. doi: 10.1073/pnas.0701644104

Woodward, J. (2009). "Scientific explanation," in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta.

Yoshimi, J. (2010). Phenomenology and connectionism. *Front. Psychol.* 2:288. doi: 10.3389/fpsyg.2011.00288

Yoshimi, J. (2012). Supervenience, dynamical systems theory, and non-reductive physicalism. *Br. J. Philos. Sci.* 63, 373–398. doi: 10.1093/bjps/axr019

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Reading, MA: Addison-Wesley.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 January 2014; accepted: 30 March 2014; published online: 23 April 2014.

Citation: Abney DH, Dale R, Yoshimi J, Kello CT, Tylén K and Fusaroli R (2014) Joint perceptual decision-making: a case study in explanatory pluralism. Front. Psychol. ***5****:330. doi: 10.3389/fpsyg.2014.00330*

This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

REVIEW ARTICLE
published: 29 April 2014
doi: 10.3389/fpsyg.2014.00373

Levels and kinds of explanation: lessons from neuropsychiatry

Sam Wilkinson *

Department of Philosophy, Durham University, Durham, UK

Edited by:
Roberto Cordeschi, Sapienza University of Rome, Italy

Reviewed by:
Lisa Bortolotti, University of Birmingham, UK
Felipe De Brigard, Duke University, USA

****Correspondence:***
Sam Wilkinson, Department of Philosophy, Durham University, Stockton Rd, County Durham DH1, 50 Old Elvet, DH1 3HN, Durham, UK
e-mail: sam.l.j.wilkinson@gmail.com

I use an example from neuropsychiatry, namely delusional misidentification, to show a distinction between levels of explanation and kinds of explanation. Building on a pragmatic view of explanation, different kinds of explanation arise because we have different kinds of explanatory concerns. One important kind of explanatory concern involves asking a certain kind of "why" question. Answering such questions provides a personal explanation, namely, renders intelligible the beliefs and actions of other persons. I use contrasting theories of delusional misidentification to highlight how different facts about the phenomenon that is being explained impose constraints on the availability of personal explanation.

Keywords: explanation in psychology, delusion, personal explanation, neuropsychiatry, levels of explanation

INTRODUCTION

Neuropsychiatry involves the study of people with mental illnesses in a way that makes use of the tools and understanding of the cognitive and brain sciences. As a result, a foundational question for neuropsychiatry is: What is the nature and extent of the contribution that the cognitive and brain sciences can make to our understanding of psychopathologies and mentally ill individuals? That is why, in this paper, I use neuropsychiatry to draw attention to two explanatory constraints. Although these constraints are important in all areas of the cognitive sciences, broadly construed, they are particularly visible in neuropsychiatry. One constraint concerns *levels* of explanation. The other constraint, which is often overlooked, often misunderstood, and is of particular importance to neuropsychiatry, concerns *kinds* of explanation.

I proceed as follows. I start by contrasting three general views about the nature of explanation, and opt for a pragmatic view. I then introduce and characterize both the "levels" and the "kinds" constraint within a pragmatic framework. I then illustrate the latter constraint by examining recent work on delusion. I end by addressing an illustrative objection.

THREE CONTRASTING VIEWS OF EXPLANATION

Before looking at explanatory constraints, it is important to reflect on what explanation is generally. Differing answers to the following two questions yield different views about the nature of explanation. These two questions are:

(1) What kinds of things are the relata in explanations? (viz. When we say that x explains y, what kinds of things are the values for x and y? Or alternatively, what kinds of things are the *explanans* and the *explananda.*)
(2) What is it for x to successfully explain y?

Following Faye (2007), I think it is useful to distinguish between three kinds of views of explanation, namely: *Formal-logical*, *Ontological*, and *Pragmatic* views of explanation. My aim is not to adjudicate between these, but rather to show that the pragmatic view provides an especially helpful way of approaching the issues in this paper.

THE FORMAL-LOGICAL VIEW

On the formal-logical view, first and famously put forward by Hempel (1965), an explanation is an abstract entity; in particular, it is a logically valid argument with propositional structure. Indeed, an *explanandum*, according to Hempel, is a proposition that follows *deductively* from an *explanans*. A number of things should be noted about this approach.

(i) Scientific and ordinary (everyday) explanations are profoundly different in nature. The things we call "explanations" in daily life never, or at best rarely, pick out logically related propositions.
(ii) This characterization is prescriptive rather than descriptive. It is neither interested in capturing how we use the word "explain," nor in capturing what scientists are actually engaged in doing when they explain things. It aims to tell us what something *ought* to be if it is to count as an explanation in this refined, ideal, sense. (One might alternatively put this in evaluative rather than constitutive terms and say that explanations are *good* explanations to the extent that they approximate this ideal.)
(iii) Explanations are objectively "out there" to be discovered.

The formal-logical view of explanation includes a number of views of explanation besides Hempel's original covering-law version. For example, it includes Salmon's statistical-relevance model as well as the unificationist theory of scientific explanation as elaborated by Friedman (1974) and Kitcher (1989).

In answer to questions 1 and 2, sets of propositions explain other propositions, and they do so by standing in valid and sound deductive relations to each other. Practically speaking, although it may apply to areas of physics, it is too demanding to usefully apply to psychology, and it does not reflect what psychologists actually do or ought to do. Of course, given the aim of this view of explanation, that is not necessarily a criticism.

THE ONTOLOGICAL VIEW

On the ontological view, explanations are not made up of logically related propositions. They are made of concrete entities like, for example, objects, states of affairs, or events. For example, you might think that events explain other events. In particular, it is common within this approach, to think of causes explaining their effects. An instance of fire explains an instance of smoke.

So, in answer to questions 1 and 2 above, we get: Events (or states of affairs) explain other events (or states of affairs), and they do so by standing in predicable law-like causal relations.

A couple of things should be noted about this view:

(i) Again, scientific and ordinary (everyday) explanations are different in nature. We rarely explain things to each other by picking out law-like causal relations.
(ii) Again, explanations are out there to be unearthed. You discover them. You find a particular event, and you unearth the explanation of that event, namely, its cause or causes.

One recent theorist, who buys into this account in philosophy of science generally, is Woodward (2003). Another, who applies a related view specifically to psychological explanation, is Donald Davidson. To simplify somewhat, Davidson (1970) takes causal *relata* to be not objects, not properties, but events, namely, he takes events to cause other events. He also takes explanations to require the picking out of a cause (which is an event) to explain an effect (which is also an event). However, these events are only explanatory "under a certain description." In other words, he is sensitive to the fact that picking out events that are causally related is not sufficient to be explanatory: you have the pick them out in a *causally relevant* way. For example, to explain why the scales go down when weighing some plums, you appeal to the weight of the plums, not their color, even though those are two aspects of one and the same event (namely, the putting of purple plums on the scales). This has some affinities with the pragmatic view. However, we will see that, crucially, the pragmatic view opens up the possibility of non-causal explanation.

THE PRAGMATIC VIEW

According to a pragmatic view of explanation, an explanation is a good answer (and, we shall see, a variety of factors, both psychological and objective, may contribute to this "goodness") to an explanation-demanding question. The relata of explanations are not events, nor are they propositions; they are speech acts that are heavily dependent on a number of contextual factors. The relevant contextual factors can include a number of things (for example, conversational context) but the most important for our purposes are the explanatory concerns of the demander of the explanation (which I will henceforth call "the demander"). An explanation has to address the explanatory concerns of the demander, and has to be (at least a candidate for being) considered satisfying. This potential subjective satisfaction is a necessary but not a sufficient condition of something being a good explanation. Obviously, there are many objectively bad explanations that we may wrongly consider satisfying (e.g., "just-so stories"). So they have to be satisfying in a non-illusory way. Different theories will flesh out what it is for something to be "satisfying in a non-illusory way," but, very roughly, it will mean that it is true or accurate, which can then be cashed out in terms of corresponding to reality, or something weaker such as "usefulness," or "assertability." The finer details of this objective criterion are not as important for our purposes (viz. of distinguishing the pragmatic view and introducing explanatory constraints that are grounded in it) as the subjective criteria. These are that the demander has to understand the candidate explanation, and that the explanation has to address the demander's explanatory concerns.

Crucially, an explanation that is objectively good by the standards of either the ontological or formal-logical view, but which leaves the demander completely in the dark, is not considered a good explanation on the pragmatic view. Explanations are relative to a particular instance of a question being asked, and have to cater to the demander's epistemic state. The demander, it must be noted, is not necessarily an individual, but could be a collective. The "question" could be asked implicitly by the scientific community as a whole (or a subset of that community), or explicitly by an individual.

There are some varieties of the pragmatic view. The view was first introduced by Van Fraassen (1980). Achinstein (1983) has an attractive version that relies heavily on the tenets of ordinary language philosophy, and Faye (2007) puts forward his own refinements. Here is what all versions of the pragmatic view have in common, in particular, in contrast to the formal-logical and the ontological views characterized above.

(i) Scientific and ordinary explanation is essentially the same. The former simply has a more regimented context (viz. the explanatory concerns are regimented and shared across a community, namely the scientific community).
(ii) Explanations, being the products of communicative acts, are not discovered as pre-formed entities. They are answerable to how things stand in the world, but they need to be selective and carefully formulated so as to be comprehensible to the demander of the explanation. In sharp contrast to both formal-logical and ontological views, explanations simply do not exist in a possible world devoid of inquiring beings that demand and give explanations. Furthermore, these explanations are demanded within a wider pragmatic context, whether it is everyday life, the court of law, the lab, or the clinic.

So, to sum up, in answer to questions 1 and 2, explanations are communicative speech acts, and they explain in virtue of satisfying the demander's explanatory concerns in a non-illusory manner (where "illusion" can arise at the level of truth or accuracy, or at the level of comprehension, namely, thinking that one comprehends when one does not). The epistemic or informational state of the demander of the explanation will in part determine her explanatory concerns, and her explanatory concerns will dictate the kind of explanation that would be satisfying. A broadly pragmatic view of explanation is what I will be building on for the rest of this paper.

DIFFERENT EXPLANATORY CONCERNS ABOUT THE SAME PHENOMENON

On a pragmatic view (in contrast to an ontological view), one phenomenon can arouse different explanatory concerns, each of which demand different explanations. Sometimes we have different explanatory concerns because we happen to be interested in different things. At other times, however, the phenomenon itself can impose constraints on what explanatory concerns are suitable, namely, what questions one should ask.

Suppose there is a plane crash. One can, for example, ask for an explanation of the plane crash in terms of poor decision-making, or neglected obligation. Or one may ask for an explanation in terms of technical problems with the plane; or in terms of the weather conditions. Depending on certain facts about the crash, either of these explanations may be unavailable. For example, if the weather had been so extreme that even the most skilled of pilots would have been unable to avoid a crash, then the explanation in terms of poor decision-making is unavailable. Conversely, if the pilot had been a terrorist who had deliberately crashed the aircraft, then an explanation in terms of the weather or the aircraft malfunctioning will be unavailable. Realizing the unavailability of certain explanations is extremely important since it should prompt us to not ask questions that have no answers.

The crucial point is this. Certain facts about the phenomenon that you are trying to understand can restrict what questions can be asked, what explanatory concerns you should have, what explanations will be available. Somebody looking to hold someone accountable for a plane crash in situations where the weather was too extreme, is asking the wrong question. This issue of asking the right questions is extremely important in all areas of science, and especially visible when looking at mental illness.

LEVELS OF EXPLANATION

The notion of "levels of explanation" is usually introduced without any particular commitment to a general view of explanation. However, it makes good sense to view it within a pragmatic framework. Indeed, the classic mention of "levels of explanation," in David Marr's book *Vision* (1982), is phrased in terms of answering three kinds of *question*:

> *Computational theory*: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
> *Representation and algorithm*: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
> *Hardware implementation*: How can the representation and algorithm be realized physically? (p. 25)

Marr's three levels makes a point that applies to any talk of levels of explanation. It is the *functionalist* point of there being multiple realizability of high-level or functional properties in lower-level properties (e.g., the property of being a bottle-opener can be physically realized in a number of different ways, as long as it opens bottles). If our explanatory concerns are about the higher-level properties (e.g., computational properties) then addressing them by drawing attention to lower-level properties (e.g., hardware properties) will be unsuitable (nevertheless, lower level implementational properties clearly impose constraints on higher level properties: you cannot make a bottle opener out of cream cheese). If one has explanatory concerns that operate at a certain level, addressing them at a different level is at best, sub-optimal, and at worst, completely irrelevant or opaque.

Some theorists see what is called the "personal level" as just another level in this sense: as a particular functional level where we are talking about whole persons, what they believe, desire, feel etc. These "personal level" properties are (if we assume physicalism) physically implemented, but they could in principle be implemented by different physical states. Dennett's doctrine of the "intentional stance" seems to view things in this way. He presents us with the following thought experiment. Suppose:

> "some beings of vastly superior intelligence—from Mars, let us say—were to descend upon us [. . .] suppose, that is, that they did not need the intentional stance—or even the design stance—to predict our behavior in all its detail" (Dennett, 1981, p. 68).

The question then is: do these Martians miss out on anything in failing to use the intentional stance, the personal-level vocabulary of beliefs, desires etc.? According to Dennett, although they might be able to predict the exact motions of the fingers and the vibrations of vocal cords during an instance of a stockbroker buying shares in General Motors, if they fail to see

> "that indefinitely many *different* patterns of finger motions and vocal cord vibrations—even the motions of indefinitely many different individuals—could have been substituted for the actual particulars without perturbing the market, then they would have failed to see a real pattern in the world they are observing" (1981, p. 69).

Note that even here, with its non-reductive take-home message, Dennett calls this "a *predictive* strategy." The plan is to predict how a causal system will behave at the relevant fineness of grain. The finger motions are not a relevant fineness of grain for gaining a predictive understanding of the stock market. The intentional stance is the relevant fineness of grain for gaining a predictive understanding of persons.

If this is correct, then the distinction between levels and kinds of explanation that I want to put forward is unnecessary. There are only levels, and one level (perhaps the "top" level) is the "personal level." Many theorists seem to subscribe to this view (which is somewhat encouraged by the presence of the word "level" in "personal level"). They take the challenge of connecting, say, a neurobiological story to a cognitive story to be the same kind of challenge as that of connecting a subpersonal and a personal story. I hope to show now that this is not the case.

KINDS OF EXPLANATION

The pragmatic view of explanation allows there to be as many kinds of explanation as there are kinds of explanatory concerns (or, which comes to the same thing, kinds of questions worth asking). And, crucially, we have explanatory concerns that are not just causal or mechanistic. We ask questions like, "Why did this person do that?," "Why does this person believe that?" We trade on the fact that there are correct and incorrect answers to these questions, and that these answers genuinely inform us. However, they do not do so by giving us a causal, predictive, understanding of the situation. These questions demand personal explanations, and personal explanations are not merely another "level" of explanation, but a different *kind* of explanation.

We can illustrate the difference between the personal and subpersonal kinds of explanation, by reflecting on different kinds of explanation-demanding question. Roughly, whereas subpersonal explanation is mechanistic or causal, answering "How" and "How come" questions (respectively), personal explanation answers a "Why" question, where "Why" is understood in a certain way.

CAUSAL AND MECHANISTIC EXPLANATIONS

Within a pragmatic view one can distinguish causal and mechanistic explanation in the following way. A causal explanation merely tells you what causes what, whereas a mechanistic explanations is far more informative in telling how a certain phenomenon comes about. A causal explanation provides some degree of understanding, and a sufficient degree of understanding for some situations, for example, if one wants to avoid a particular effect. Thus we might establish that smoking causes cancer. We might not know exactly how it does so, but knowing that, at least, is enough to suggest that (*ceteris paribus*), if we do not want cancer, we should not smoke. We can think of causal explanations as answering a certain kind of question, namely, a "How come?" question. "How come he got cancer?" This is often expressed with a causal use of "Why," as in "Why did he get cancer?" This causal use of "why" is very different from a justificatory use that we are about to encounter.

A mechanistic explanation answers instead a "How?" question. It provides not just the cause, but the mechanism whereby a certain causal process operates. Using the cancer example, it is not enough to know that smoking causes cancer: what is required is a description of the mechanism, for example, in terms of carcinogenic disruption of genetic material through radiation given off by substances present in tobacco smoke. Mechanistic explanation provides a greater degree of understanding than merely citing causes. It is reasonable to think of mechanistic and causal explanations as being the same kind of explanation insofar as the kinds of concerns addressed are of the same kind; namely, of predicting how a brute causal system will behave in relevant counterfactual circumstances.

PERSONAL EXPLANATIONS

As I said, we do sometimes use "why," when our explanatory concerns are causal or mechanistic, for example, when we ask "Why is there a hole is the ozone layer?" We mean, "By what cause or process is there a hole in the ozone layer?" We know it is not there for a *reason*. However, when we ask, "Why is there a STOP sign at the end of that road?" we are asking for a reason, a justification, a *rationale*, for its being there. Along with the distinction between justificatory and causal uses of "why" in the *question* ("Why is there a STOP sign there?/Why did you raise your hand?" vs. "Why did he get cancer?") we have the distinction between justificatory and causal uses of "because" in the *answer* ("Because there tends to be fast-moving traffic in the main road/Because I wanted to ask a question" vs. "Because he smoked too heavily"). Answering such a "why" question involves citing a person's reasons or grounds for believing certain things and acting in certain ways (or the general, publicly agreed, reasons, not attributable to a specific person, as in the case with the STOP sign).

With beliefs and actions, we often ask questions of one another: "Why do you believe that?" "Why did you do that?" In doing this, we are asking a very particular kind of question, and one that requires a very particular kind of answer. This answer is commonly called a *rational* explanation (Davidson, 1963). However, "rational" has a categorical and evaluative sense. The opposite of "rational" in the categorical sense is "a-rational" (or "non-rational"), whereas the opposite of "rational" in the evaluative sense is "irrational." The way "rational" is used here is categorical, not evaluative. As we are about to abundantly see, you can have rational explanations of *irrational* phenomena. Indeed, something that cannot be given a rational explanation cannot be irrational; it is merely *a*-rational. Consider a nervous tick. You cannot ask why (in the justificatory sense) someone with a nervous tick is behaving the way they are. And, clearly, you cannot evaluate their reasons as bad reasons if there are none. However, because of this ambiguity with the word "rational" I have made the terminological decision to use "personal explanation" rather than "rational explanation."

If you ask me, "Why did you raise your hand?" and I answer, "Because I wanted to ask a question," that is normally a satisfying explanation. If I tell you a full physiological story about what happened up until the point when my hand went up, that may be interesting, but it is not an answer to *that* question. You were after a *reason*. You wanted to know what I was hoping to achieve by raising my hand. The same applies when you ask the question "Why do you believe this?" You are after *reasons* for my belief, not any mechanistic story. You want to know what grounds I have, if any, for believing something.

IS PERSONAL EXPLANATION CAUSAL EXPLANATION?

Now, you might agree that these are common and valid explanatory practices. However, you might question whether there is anything fundamentally different about them. Why are these not causal explanations? We know that certain beliefs and desires in certain contexts will give rise to certain actions. This seems rather causal.

It is worth noting that there are those who are fully accepting of the explanatory autonomy of reason-giving explanation, but who also claim that reasons are causes. For example (unlike, say, Anscombe, 1957 and the early Dennett, 1969) Davidson sees reasons as causes. Now, although I prefer not to think of reasons as causes, I am willing to accept for the sake of argument, that reasons are causes, especially if one accepts a counterfactual theory of causation. A counterfactual theory of causation is (roughly) one whereby A causes B, if and only if, had there not been A, there would not have been B. If I had not wanted to ask a question, I would not have raised my hand. In that very simple sense, my *desire* to ask a question was a cause of my *action*. However, it does not follow from something being a cause, that something is explanatory *in virtue of being a cause*. Since our explanatory concerns when we ask this special variety of "why" question are not causal concerns, the explanation given in terms of reasons is not a *causal explanation*, even if one thinks that reasons are causes. In short, conceding that reasons are causes does not concede that rational explanations (explanations in terms of reasons) are causal explanations.

So, what are our explanatory concerns when we ask for a personal explanation and why are they not causal? We want to understand the person *qua* agent, not *qua* causal system. Suppose somebody behaves in a way that, unlike a nervous tick, looks controlled and deliberate, and yet you still cannot give it an explanation. To take an unrealistic example, suppose someone holds up a poisonous mushroom, and announces: "This mushroom is deadly poisonous and I have no intention of killing myself" but then pops the mushroom in his mouth. This behavior is *surprising*, but it is not just surprise that you feel, but perplexity and confusion. In particular, you do not understand *why* this person ate the mushroom. You can hypothesize that he was lying, either about the poisonous nature of the mushroom, or his intentions to stay alive. Or he was demonstrating an antidote. Or he doesn't know the meaning of one or more of the words he was using. However, taken at face value, this action is perplexing. When this happens, we are not just bemoaning a failure to predict. There is more to it than this: we are perplexed by this *person qua agent*, by the fact that we find the person *unintelligible*. In fact, it seems that if his behavior can in no way be reinterpreted so as to confer intelligibility then the best way to understand it is as a brute causal process (perhaps he's a realistic-looking android, programmed to behave in just this way to prove the point I'm making here). But explaining it in these terms would require us to ask (and answer) a different question ("How come?" rather than "Why?").

Other human beings often behave in ways that we have failed to anticipate, but that are still perfectly intelligible. We might say that, while causal and mechanistic explanations confer *predictability*, personal explanations confer *intelligibility*. Of course, assuming that people will be intelligible makes them more predictable, in the sense that it narrows down the ways they might behave in certain circumstances, but that does not mean that any given personal explanation improves, or is aimed at improving, our causal understanding of a person. Indeed, all of the causal understanding we need is already in place: we know that certain beliefs and desires give rise to certain actions, certain kinds of evidence give rise to certain beliefs. We just want to know, in this instance, what beliefs, or desires, or evidence the subject actually had, so that we can understand and evaluate them as persons.

ON THE RELATIONSHIP BETWEEN SUBPERSONAL AND PERSONAL EXPLANATIONS

We have looked at what personal explanation is, and what causal and mechanistic subpersonal explanations are. But what is the relationship between them? We will start by looking at how they can compete, and then we will look at how they can inform each other.

HOW THEY COMPETE WITH EACH OTHER

Personal and subpersonal explanations do not compete with each other in the way that they compete amongst themselves. Within a pragmatic framework competing explanations are competing answers to the *same question*. When you compare a personal and a subpersonal explanation, you are comparing answers to *different questions* (and indeed to different *kinds* of questions). However, competition comes in at a different level: at the level of asking the question in the first place. Asking a certain question presupposes that it is appropriate, that it can be answered, and that presupposes certain facts about the phenomenon in question. So, personal and subpersonal explanations will not directly compete (in the way that, as we are about to see, two personal explanations *can* directly compete). However, in some cases, both of them *being offered at all* will presuppose the obtaining of two incompatible states of affairs. Two demanders of explanations for a plane crash might ask: "Who was to blame for the crash?" or "What kind of weather conditions caused the crash?" Each implies different facts concerning the plane crash (e.g., the former implies that blame is to be attributed and that it wasn't a so-called "Act of God"). Asking a question betrays assumptions about the phenomenon that you are asking questions about. We will see more on this when we look at examples from delusion.

Another important way in which personal and subpersonal explanations compete is by imposing constraints on one another. If a personal explanation claims that, for example, the subject believed that *p* because they had an experience with a certain quality, but the best available subpersonal account suggests that the experience could not have been like that, then clearly these two explanations conflict. However, we will see that it is precisely through this constraining that the two kinds of explanation can inform each other.

HOW THEY INFORM EACH OTHER

We can use both personal and subpersonal explanations to further our understanding of the same individual. To use an example that will be relevant to us, we can ask about a patient with the Capgras delusion the subpersonal question, "How has this brain damage disrupted normal cognitive functioning?" A really good answer to this will make it altogether unmysterious why (how come) this particular damage has disrupted functioning in this particular way, and not in any other way. This will require causal and mechanistic explanations at different levels. However, one can also ask, "Why (on what grounds) does she believe what she does?" In answering this, you cannot use the same vocabulary as when answering the first, subpersonal, question. Dopamine dysregulation, modular damage, etc. none of these are even the right kind of thing to provide grounds for the subject. Similarly, to take a non-pathological case, you might ask me:

Q: Why did you think that James was at home?
And I might answer:
A: Because I saw his car in the driveway.

You would think it some kind of joke if I instead gave you a story about light hitting my retina, causing activation in V1, and so on. You want to know on what grounds I came to believe what I did.

Although subpersonal vocabulary (neurotransmitters, processing streams and so on) cannot feature directly in personal explanation, this is not to say that subpersonal psychology (broadly construed to include all the cognitive and brain sciences) cannot make very important contributions to personal explanations. In particular, it can make two very different kinds of contribution.

First, it can give us an idea of the nature of the grounds that a subject might have (e.g., what experiences or emotions they might be undergoing) and how it is that they have them, or rather, how come they have these experiences and not others. For example, as we are about to see, subpersonal psychology can suggest that the Capgras patient is experiencing a feeling of unfamiliarity toward a loved one. Once we understand what the subject may be experiencing, there is scope for their beliefs to be rendered *intelligible*, namely, to be subject to personal explanation. We can answer the question: "*Why* does the subject believe this?" In other words, the first kind of contribution that subpersonal psychology can make to personal explanation is one of suggesting the starting point for such an explanation.

The second kind of contribution that subpersonal psychology can make is very different. It may be able to warn us when personal explanation is *unavailable*. That is to say, it may warn us when any attempts at understanding the subject in terms of subjective grounds would be a waste of time. We saw with our plane crash analogy that an understanding of the situation may lead us to conclude that, for example, no blame is to be attributed. Similarly, an understanding of certain mentally ill individuals may lead us to realize that there is no answer to the "why" questions, "Why did he do this?," "Why does he believe this?" Sometimes there may simply be *no grounds* for certain beliefs and behaviors, and it is vital that subpersonal psychology can warn us when this may be the case.

AN EXAMPLE FROM NEUROPSYCHIATRY: DELUSION

Opposing views about the nature of delusional misidentification map on, in a nicely illustrative manner, to these two kinds of contributions that subpersonal psychology can make to personal explanation.

A key figure in the history of theoretical work on delusion is Karl Jaspers, who, in his *General Psychopathology* (1963), claimed that there were two very different projects: one of "understanding the subject," and the other of rendering the psychopathological phenomenon causally tractable. The first project roughly corresponds to personal explanation, whereas the latter corresponds to subpersonal explanation. He thought that delusions (in particular the primary delusions of schizophrenia) arise without any grounds or justification; they are "un-understandable" in the sense that they are not intelligible. We cannot answer the question: "Why (viz. on what grounds) do these subjects believe what they do?" All we can do is try to understand the subject *qua* causal system.

However, half a century later the way was paved for (at least some delusions) to be rendered intelligible. In particular, Brendan Maher presciently hypothesized that the "delusional belief is not being held "in the face of evidence strong enough to destroy it," but is being held because evidence is strong enough to support it" (1974, p. 99). What we then have to do as theorists is figure out what the evidence is, and how it arises. This will obviously have the potential to vary from one delusion to another, and may indeed provide a satisfying explanation of why some patients have some delusions and others have others (viz. there will be an unmysterious connection between the nature of their experience, and the content of their delusion).

THE CLASSIC BOTTOM-UP MODEL

This project received something of a breakthrough (a full 16 years later) in the case of the Capgras delusion (the delusion that one or more loved ones have been replaced by identical-looking impostors). Borrowing Bauer's (1984) model for facial processing, whereby there are two streams for processing facial information—one covert, affective and anatomically dorsal, the other overt, semantic, and anatomically ventral—Ellis and Young (1990) put forward the influential proposal that the Capgras delusion can be understood as a sort of "inverse prosopagnosia."

People with prosopagnosia have difficulty in the overt recognition of faces. Show them a picture of a familiar face and they will not be able to tell you whose face it is. And yet, surprisingly, some of them appear to have differential autonomic responses (roughly, affective/emotional responses) to these faces, as measured by heightened skin conductance response (SCR). In other words, although they themselves cannot tell you whose face they are looking at, their affective system seems at the very least to be able to "tell" that it is someone familiar. Ellis and Young hypothesized that Bauer's two streams can be selectively impaired, leading to double dissociation. According to them, whereas with prosopagnosia the affective stream for "covert recognition" is intact and the semantic stream for "overt recognition" is impaired, with the Capgras delusion it is the other way around. This means that the Capgras patient is presented with someone who, thanks to intact semantic processing, looks to them exactly like a loved one, but there is a lack of affective response. The perceived person feels unfamiliar and the patient therefore concludes that this person cannot be the loved one in question. This model was given experimental support (Ellis et al., 1997) when it was discovered that, in contra-distinction to prosopagnosia, Capgras patients show diminished SCR when presented with familiar faces.

According to this etiology, there is scope for the Capgras delusion to be rendered intelligible, to be given a personal explanation, since it can be seen as something that is inferred on the basis of experiential evidence. These theories, which take delusions to be grounded in unusual experiences, are called "bottom-up" theories. A complete bottom-up account will contain a mix of personal and subpersonal explanation. The very existence of the anomalous experience is explained in terms of mechanism (in the Capgras case, on the Ellis and Young model, this could involve explaining how lesions disrupt affective processing of familiar faces). But the judgment itself is personal. The *person* infers from their experience. And the relevant question to ask is: "Why does the *person* believe that this woman is not his mother?" And the relevant answer is roughly: "Because this woman feels deeply unfamiliar to him." This is not a causal, mechanistic explanation, but a personal one. And, if correct, it tells us all we need to know *within the scope of personal explanations*.

EXPLANATIONIST vs. ENDORSEMENT ACCOUNTS

However, there is a debate within bottom-up theories about what *precisely* the correct answer to this question (viz. "Why does this person believe that this woman is not his mother?") is. Phrased in more technical terms, there is a debate about the content of the experience in the Capgras delusion. In other words, what exactly does the subject's experience tell her; how does it subjectively support her judgment?

To borrow Bayne and Pacherie's (2004) terminology, "explanationist" accounts (e.g., Ellis and Young, 1990, Maher, 1999) claim that the content of the Capgras patient's experience is something *sparse* like, "This woman feels strange," and that the delusional judgment *explains* the bizarre experience (roughly, the subject reasons: "This woman, in spite of looking like my mother, doesn't *feel* like my mother would feel, therefore she cannot *be* my mother"). The opposing accounts, so-called "endorsement" accounts (e.g., Bayne and Pacherie, 2004) claim that the delusional content is encoded directly in the unusual experience, and all that suffices is endorsement of that content. The content of the Capgras patient's experience, on such a view, is something *rich* like, "This woman is not my mother."

Now, bracketing the plausibility of either account, it is worth noting an explanatory trade-off. As Pacherie (2009) points out, the explanationist account can more easily explain how the experience gets its content, since the content is so sparse. It can simply say that there is a disruption in emotional or affective processing. The task for subpersonal psychology is comparatively easy. However, it is a bigger explanatory step from the sparse content of the experience to the rich content of the delusion. One *prima facie* problem with this is that, if the experience is sparse and nonspecific, why is there not a wider array of potential hypotheses used to explain it? ("Maybe I don't like mum anymore," "Maybe I'm tired" etc.). In contrast, the endorsement theorist can get from the experience to the judgment just fine, since they have the same content. However, as Pacherie puts it, where "the endorsement account would appear to be weakest is in explaining how delusional patients could have the experiences that the account says they do" (Pacherie, 2009, p. 107). More precisely, subpersonal psychology needs to step in and tell us how it is that an experience can have a rich content like, "This woman is not my mother."

Here we get a nice illustration of two directly competing personal explanations, namely, different answers to the very same "why" question. It also illustrates how these give rise to two different explanatory burdens that need to be picked up by subpersonal explanations. Presenting the competing accounts in terms of questions, where "Why?" and "How come?" questions correspond to the personal (justificatory) and subpersonal (mechanistic) questions respectively, we get (roughly) the following. For the explanationist account we get:

> Q: "Why does the subject believe that his mother has been replaced by an impostor?"
> A: "Because she feels unfamiliar to him."
> Q: "How come she feels unfamiliar to him?"
> A: "Because affective processing has been disrupted in such and such a way."

For the endorsement account we get:

> Q: "Why does the subject believe that his mother has been replaced by an impostor?"
> A: "Because his experience presents this woman as not being his mother."
> Q: "How come?"
> A: "Because (for example) subpersonal mechanisms responsible for managing the representation of the identities of known individuals has been disrupted" (see e.g., Wilkinson, in press).

BACK TO UNINTELLIGIBILITY: TOP-DOWN ACCOUNTS

However, not everyone subscribes to bottom-up theories of delusions (Eilan, 2000; Campbell, 2001). In a way that harks back somewhat to Jaspers, these theorists claim that the delusion is not inferred, nor grounded in evidence, but caused. Any report (or even experimental evidence from SCR), for example, that the mother feels unfamiliar, is a consequence of (or an accompaniment to) the delusional belief, but not grounds for it. She feels unfamiliar because she is judged to not be the subject's mother, and not the other way around. As Campbell puts it, "'delusion' is a matter of top-down disturbance in some fundamental beliefs of the subject which may consequently affect experiences and actions" (2001, p. 89). An upshot of this is that the belief can only be explained subpersonally, and, of course, this leaves a large explanatory burden for subpersonal psychology. We cannot answer the question "*Why* does the *person* believe that this woman is not his mother?" We cannot appeal to grounds since there are none. We are back to Jaspers' claim that delusional subjects are "un-understandable." The only question with an illuminating answer is: "What has *caused* this person to believe what she does?" This is precisely what I meant when I said that some etiologies will take personal explanation to not be available. However, note that, although, on these top-down theories, the delusional belief may not be amenable to personal explanation, any *action* performed on the basis of the belief will be, and this explanation will appeal to the belief. In such a situation we roughly get the following series of questions and answers.

> Q: "Why did the patient stab her father (even though they seemed to have a good relationship prior to the event)?"
> A: "Because she believed that he was not her father, but an identical-looking impostor."
> Q: "And on what grounds did she believe this?"
> A: "There were none. The belief was merely caused."

At this point we would need to delve into the subpersonal mechanisms to understand what is underpinning the (groundless) belief in terms of mechanisms.

AN ILLUSTRATIVE OBJECTION

Somebody might say that the "kinds of explanation" constraint is illusory, and, in particular, that personal explanation has no place in a scientific enterprise. When people believe things on epistemic grounds, or they do things for reasons, nothing fundamentally different is going on. This could, in principle, all be explained subpersonally by the cognitive sciences.

I think this is a misleading way to think, and I would like to run through a thought experiment to illustrate why. Suppose we had a "complete" subpersonal description (whatever that means!) of what is going on, say, in an instance of thought insertion. We are not appealing to anything personal, we are not talking about grounds or reasons, just mechanistic stuff. Assuming physicalism, and highly advanced imaging techniques at our disposal, we could, in principle, take any given individual and see when they would (and when they would not) report inserted thoughts. I am very happy to grant this. Thought insertion is, in a rather trivial sense, a fundamentally physical phenomenon. However, suppose that it happens that thought insertion, the denial of ownership of one's thoughts, is actually grounded in a very bizarre experience. We *could* know exactly what is going on in the brain of someone (i.e., we know what that neural activity looks like on the scanner) who is reporting inserted thoughts, but still not know on what grounds they are denying ownership of their thoughts (or indeed that there are grounds at all). This seems like a possible epistemic state for us, as scientists, to be in. More worryingly, it entails ignorance of an important and irreducible fact (by "fact," I mean, a claim that is determinately true) namely, concerning the grounds on which somebody is denying ownership of her thoughts.

As the section in this paper that examines the relationship between personal and subpersonal explanations makes clear, I am not denying that we could work out, to some extent, the subject's experiential grounds from subpersonal data. Here, I have advocated the fruitful, but careful, contribution of subpersonal psychology to personal explanation. But this in itself requires us to take personal explanation seriously. In this thought experiment, personal explanation is disregarded, eliminated, out of bounds. And the intuition I hope you share is that this entails a kind of ignorance.

Of course, where this illustrative objection goes awry is in making the fallacious step from physicalism to reductionism. Everything in the universe could well be made of physical stuff, but as human beings we are constrained, partly by our interests and concerns, and partly by our cognitive and conceptual limitations. We therefore talk about, and explain, many different things (and kinds of things) in many different ways (and kinds of ways). Obviously, there are facts concerning people that go beyond, and are not reducible to, neural facts. There are epistemic facts (facts about grounds for belief), motivational facts (facts about people doing things for reasons). There are also facts about what people experience (it is, for example, a fact, as unambiguously true as anything, that I am currently not in excruciating pain). And it does not stop there: there are social facts, economic facts, contractual facts, and so on. You cannot capture what it is for me to sign a contract using physics and physiology.

SUMMARY AND CONCLUSION

In this paper, I built on a pragmatic view of explanation, on the basis of which explanations are answers to explanation-demanding questions, in order to show a distinction between levels of explanation and kinds of explanation. Different kinds of questions, and hence explanations, arise because we have different kinds of explanatory concerns. One important kind of explanatory concern involves asking a certain justificatory kind of "why" question. Answering this kind of question provides a personal explanation, namely, renders intelligible the beliefs and actions of other persons. I then used contrasting theories of delusional misidentification to illustrate how different facts about the phenomenon that is being explained impose (i) constraints on the availability of personal explanation and (ii) leave different explanatory burdens for subpersonal psychology (broadly construed). More generally, this also illustrated how asking certain kinds of questions, seeking certain kinds of explanations, carries implicit assumptions about the nature of the phenomenon about which the questions are being asked.

ACKNOWLEDGEMENTS

This research was supported by a Wellcome Trust Strategic Award WT098455MA.

REFERENCES

Achinstein, P. (1983). *The Nature of Explanation*. New York, NY: Oxford University Press.

Anscombe, G. E. M. (1957). *Intention.* Oxford: Basil Blackwell.

Bauer, R. M. (1984). Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the guilty knowledge test. *Neuropsychologia* 22, 457–469.

Bayne, T., and Pacherie, E. (2004). Bottom up or top down? *Philos. Psychiatry Psychol.* 11, 1–11. doi: 10.1353/ppp.2004.0033

Campbell, J. (2001). Rationality, meaning and the analysis of delusion. *Philos. Psychiatry Psychol* 8, 89–100. doi: 10.1353/ppp.2001.0004

Davidson, D. (1963). Actions, reasons and causes. *J. Philos.* 60, 685–700.

Davidson, D. (1970). "Mental events," in, *Experience and Theory*, eds L. Foster and J. W. Swanson (Amherst: University of Massachusetts Press), 79–101.

Dennett, D. C. (1969). *Content and Consciousness.* London: Routledge.

Dennett, D. C. (1981). "True believers: the intentional strategy and why it works," in *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*, ed A. F. Heath (New York, NY: Clarendon), 150–167.

Eilan, N. (2000). "On understanding schizophrenia," in *Exploring the Self*, ed D. Zahavi (Amsterdam: John Benjamins), 97–113.

Ellis, H. D., and Young, A. W. (1990). Accounting for delusional misidentifications. *Br. J. Psychiatry* 157, 239–248. doi: 10.1192/bjp.157.2.239

Ellis, H. D., Young, A. W., Quayle, A. H., and de Pauw, K. W. (1997). Reduced autonomic responses to faces in Capgras delusion. *Proc. R. Soc. Lond. Biol. Sci. B* 264, 1085–1092. doi: 10.1098/rspb.1997.0150

Faye, J. (2007). "The pragmatic-rhetorical theory of explanation," in *Rethinking Explanation. Series: Boston Studies in the Philosophy of Science*, Vol. 252. eds J. Persson, and P. Ylikoski (Dordrecht: Springer), 43–68.

Friedman, M. (1974). Explanation and scientific understanding. *J. Philos.* 71, 5–9.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York, NY: The Free Press.

Jaspers, K. (1963). *General Psychopathology.* Trans. J. Hoenig and M. Hamilton. Manchester: Manchester University Press.

Kitcher, P. (1989). "Scientific understanding and the causal structure of the world," in *Scientific Explanation*, eds P. Kitcher and W. Salmon (Minneapolis, MN: University of Minnesota Press), 410–505.

Maher, B. A. (1974). Delusional thinking and perceptual disorder. *J. Individ. Psychol.* 30, 98–113.

Maher, B. A. (1999). Anomalous experience in everyday life: its significance for psychopathology. *Monist* 82, 547–570.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Pacherie, E. (2009). "Perception, emotions and delusions: revisiting the capgras delusion," in *Delusion and Self Deception: Affective and Motivational Influences on Belief Formation*,eds T. Bayne and J. Fernàndez (Hove: Psychology Press), 105–123.

Van Fraassen, B. C. (1980). *The Scientific Image*. New York, NY: Oxford University Press.

Wilkinson, S. (in press). Delusions, dreams, and the nature of identification. *Philos. Psychol.* 1–24. doi: 10.1080/09515089.2013.830351

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 08 April 2014; published online: 29 April 2014.

Citation: Wilkinson S (2014) Levels and kinds of explanation: lessons from neuropsychiatry. Front. Psychol. ***5****:373. doi: 10.3389/fpsyg.2014.00373*

This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 19 May 2014
doi: 10.3389/fpsyg.2014.00440

The subjective meaning of cognitive architecture: a Marrian analysis

Sashank Varma*

Department of Educational Psychology, University of Minnesota, Minneapolis, MN, USA

Edited by:
Roberto Cordeschi, Sapienza University of Rome, Italy

Reviewed by:
John Tracy Hale, Cornell University, USA
Kevin Gluck, Air Force Research Laboratory, USA

Correspondence:
Sashank Varma, Department of Educational Psychology, University of Minnesota, 56 East River Road, Minneapolis, MN 55455, USA
e-mail: sashank@umn.edu

Marr famously decomposed cognitive theories into three levels. Newell, Pylyshyn, and Anderson offered parallel decompositions of *cognitive architectures*, which are psychologically plausible computational formalisms for expressing computational models of cognition. These analyses focused on the *objective meaning* of each level – how it supports computational models that correspond to cognitive phenomena. This paper develops a complementary analysis of the *subjective meaning* of each level – how it helps cognitive scientists understand cognition. It then argues against calls to eliminatively reduce higher levels to lower levels, for example, in the name of parsimony. Finally, it argues that the failure to attend to the multiple meanings and levels of cognitive architecture contributes to the current, disunified state of theoretical cognitive science.

Keywords: cognitive architecture, unified theories of cognition, computational models, reduction, parsimony, identifiability

INTRODUCTION

In the first chapter of *Vision*, Marr (1982) famously decomposed cognitive theories into three levels. He examined neuroscience theories of vision and found them too focused on neural circuitry – the lowest level. He examined artificial intelligence models of vision and found them too focused on data structures and algorithms – the middle level. He argued that understanding the *what* and *how* of vision would not constitute a complete theoretical account. An understanding of the *why* of vision – the problem it solves for the organism – was also needed, and this could only be provided by the highest level. Marr surveyed the cognitive science landscape and found only two theories articulated at this level, Chomsky's (1965) theory of language "competence" and Gibson's (1979) "ecological" theory of visual perception. He argued that future progress in cognitive science would require greater attention to all three levels.

Marr was not the only cognitive scientist thinking along these lines. The cognitive revolution was 25 years old when *Vision* was published, and the optimism generated by early computational models was giving way to a growing recognition of their limitations. Newell (1982), Pylyshyn (1984), and Anderson (1990) offered analyses that were strikingly similar to Marr's in the levels they proposed, but that addressed a class of cognitive theories called *cognitive architectures*.

Every science strives for a unified theory of all of its phenomena (Oppenheim and Putnam, 1958). For example, physicists are searching for a grand unified theory of the fundamental forces of nature (Weinberg, 1993). Its equations, once discovered, will provide an account of all physical phenomena – at least in principle. Cognitive scientists are similarly searching for a *unified theory of cognition* (Newell, 1990). It will not be a set of equations, as it will be for physics. Rather, it will be a cognitive architecture – a computational formalism for expressing computational models of cognitive phenomena. This reflects the fundamental claim of the cognitive revolution, that cognition is a form of information processing. A better analogy, then, is to classical mechanics, the unified physical theory of its day. Classical mechanics postulates a continuous universe where forces act on bodies across space and time. Newton lacked a suitable mathematical formalism for expressing classical mechanics, and so he designed one – the calculus. Similarly, cognitive architects design new computational formalisms for expressing the models that cognitive scientists dream up.

The analyses offered by Marr, Newell, Pylyshyn, and Anderson focused on the *objective meaning* of each level – how it supports models that correspond to the phenomena of cognition. This paper offers a complementary analysis of the *subjective meaning* of each level – how it helps cognitive scientists understand cognition (Varma, 2011). The first half articulates the objective and subjective meanings of each level. The important point is that these meanings are quasi-independent: they can mutually constrain each other ("quasi"), but cannot entirely replace each other ("independence"). This paper then draws the implications of this analysis. It first argues that the subjective meanings of different levels are also quasi-independent, and this precludes the reduction of higher levels to lower levels, for example, in the name parsimony. In fact, preserving multiple levels provides working cognitive scientists with the flexibility to choose the most appropriate level for their modeling activities. It concludes by explaining the current, disunified state of theoretical cognitive science as the product of failing to understand the multiple meanings and multiple levels at which architectures explain cognition, and on which they must be compared.

THE LOWEST LEVEL

The lowest level of cognitive architecture is the most familiar; it is what cognitive scientists think of when they think of architecture at all. This section first describes the objective meaning of the lowest level, as articulated by Marr and others, and then describes its subjective meaning.

COMPUTATIONAL MECHANISMS

Standard decompositions of cognitive architecture differ in how they name the lowest level. Marr (1982) called it the "hardware implementation," Newell (1982) the "device" level, Pylyshyn (1984) the "physical" (or "biological") level, and Anderson (1990) the "biological" level. What is common to all is the proposal that the lowest level defines the interface between the brain and the mind, where neural information processing elements aggregate into cognitive information processing elements. We call these cognitive information processing elements *computational mechanisms*. They come in three types.

Basic representations are the primitive means for encoding information. Different architectures provide different basic representations. For example, the basic representation of production system architectures is the declarative memory element, or dme (Newell, 1973a). A dme is a set of attribute–value pairs, where each attribute is a distinction that the perceptual-cognitive-motor system makes and each value is a number, symbol, or another dme. The basic representation of connectionist architectures is the vector of microfeatures, where each microfeature has a numeric value, encoded as the activation level of a unit (Rumelhart et al., 1986). The basic representation of exemplar architectures is the episodic trace. It is a vector of features, some encoding semantic information and others contextual information, that can assume numeric values (Raaijmakers and Shiffrin, 1981). A critical difference between dmes on one hand and microfeature vectors and episodic traces on the other is that the former can be recursively embedded within each other, whereas the latter are "flat," and thus recursive embeddings must be implemented by a combination of computational mechanisms (Elman, 1990, 1993; Hinton, 1990; Pollack, 1990; Smolensky, 1990; Murdock, 1993).

Basic operators are the primitive means for processing information. A basic operator takes basic representations as input, transforms them, and generates basic representations as output. For example, the basic operator of production system architectures is the production (Newell, 1973a). A production has a condition side and an action side. The condition side is matched against the available dmes. If a match results and the production is fired, then the individual actions of the action side are executed, adding, deleting, and modifying dmes. The basic operators of connectionist architectures include the weighted links that connect units and the activation functions of the units themselves (Rumelhart et al., 1986). For example, in a feedforward connectionist architecture, as activation flows across weighted links and through activation functions, input vectors are transformed into hidden layer vectors, and ultimately into output vectors. The basic operator of exemplar architectures computes the similarity between two episodic traces. Similarity is a superlinear function of the number of shared feature values – multiplicative in the search of associative memory model (SAM; Raaijmakers and Shiffrin, 1981), cubic in Minerva-II (Hintzman, 1986), and exponential in the generalized context model (GCM; Nosofsky, 1984; Shepard, 1987).

The *control structure* is the regimen for scheduling the application of basic operators to basic representations over time (Newell, 1973a). Different architectures adopt different control structures. Among production system architectures, ACT-R employs serial control, firing one production at each point in time (Anderson, 2007), whereas 4CAPS employs parallel control, firing all matching productions (Just and Varma, 2007). Soar utilizes a mixed control structure, parallel for some aspects of its "decision cycle" and serial for others (Newell, 1990; Laird, 2012). Connectionist architectures also exhibit a variety of control structures: Hopfield (1982) networks update the activation of one unit at a time; interactive activation and competition (IAC) networks update all units simultaneously (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982); and feedforward networks mix the two control structures, updating units in the same layer in parallel and units in different layers serially (Rumelhart et al., 1986). Exemplar architectures offer comparatively rudimentary control structures, perhaps owing to their origins in mathematical psychology, not computer science. One exception is Minerva-II, which uses a serial control structure where the trace retrieved on the current iteration serves as the probe on the next iteration. This continues until the content of the probe and the retrieved trace converge (Hintzman, 1986).

CRITERIA FOR COMPUTATIONAL MECHANISMS

Cognitive scientists prefer to construct computational models within cognitive architectures rather than general-purpose programing languages such as C and Java. This is because the computational mechanisms of architectures are psychologically plausible (e.g., microfeature vectors), whereas those of programing languages are not (e.g., "for loops"). This decreases the degrees of freedom available during the construction of models, increasing their generalizability to new phenomena.

There are two criteria for judging the psychological plausibility of computational mechanisms. The first criterion is that computational mechanisms be *biologically realizable*. Prior analyses of the lowest level define it as the interface between the mind and the brain. Marr populated his lowest level with neural processing elements such as feature detectors (e.g., Hubel and Wiesel, 1962) and spatial frequency detectors (e.g., Campbell and Robson, 1968). However, he acknowledged the parallel between the neural architecture and "the detailed computer architecture" (Marr, 1982, p. 25). Newell (1989) offered a similarly dual conception of the lowest level, noting that in "current digital computers it is the register-transfer level, but in biological systems it is some organization of neural circuits" (p. 404). For a computational mechanism to be biologically plausible, it must be consistent with what is known about neural information processing. It has been claimed that the computational mechanisms of connectionist architectures are of greater biological realizability than those of symbolic architectures (Rumelhart and McClelland, 1986). There are two reasons to doubt this claim. The first is that some neuroscientists question the correspondence between the computational mechanisms of connectionist architectures and the details of neural information processing (Crick and Asanuma, 1986, pp. 369–371). The second reason is that the biological realizability of the computational mechanisms of some symbolic architectures has been demonstrated by the construction of models that can account for neuroscience data (Anderson, 2007; Just and Varma, 2007).

The second criterion is that computational mechanisms be *disaggregate* (Newell and Simon, 1972; Pylyshyn, 1984). A computational mechanism is disaggregate if it can be defined in non-cognitive terms. A non-cognitive definition can be mathematical, physical, chemical, or biological. By contrast, a cognitive definition is in terms of other computational mechanisms. A cognitively defined computational mechanism is problematic because if it is replaced everywhere (i.e., in all models) with its defining combination, the resulting architecture would have the same expressive power but would be more parsimonious, and would therefore be preferable. The computational mechanisms of connectionist architectures are disaggregate, and therefore do well on this criterion. Units, weighted links, activation functions, and learning rules can be defined mathematically, without recourse to cognitive terms. By contrast, the computational mechanisms of symbolic architectures are on shakier ground. For example, the basic operator of production system architectures, the production, directly supports "variable binding" (Fodor and Pylyshyn, 1988). Some connectionists have argued that variable binding is an aggregate computational mechanism, and that it should be replaced everywhere with a combination of simpler computational mechanism, for example, in the "conjunctive coding" technique (Hinton et al., 1986; Touretzky and Hinton, 1988).

COGNITIVE PRIMITIVES

The subjective meaning of a cognitive architecture is the understanding it brings cognitive scientists of cognition (Varma, 2011). At the lowest level, the computational mechanisms of an architecture are *cognitive primitives* that specify a metaphysics for cognition. They offer a particular perspective on cognitive information processing, guiding cognitive scientists to value some computational models over others that are "equivalent" in objective meaning (i.e., correspondence to cognitive phenomena).

That the lowest level makes metaphysical claims is hinted at in Marr's analysis. He observed that choices made at the lowest level necessarily make it easier to express some cognitions (i.e., more natural, more parsimonious) but harder to express others (i.e., more awkward, more complex). He illustrated this with an example from mathematics: choosing a base-10 representation for numbers makes some computations easy, such as determining whether a number is a power of 10, but makes other computations difficult, such as determining whether a number is a power of 2. If a base-2 representation is chosen, however, the opposite tradeoff results. More generally, "any particular representation makes certain information explicit at the expense of information that is pushed into the background and may be quite hard to recover" (Marr, 1982, p. 21).

Cognitive primitives are not computational mechanisms; the subjective meaning of the lowest level is quasi-independent of its objective meaning. This is evidenced by the fact that different cognitive primitives can be realized by the same computational mechanism, and the same cognitive primitive can be realized by different computational mechanisms. Consider the productions of the ACT-R and 4CAPS architectures. As computational mechanisms, they are quite similar: their condition sides are matched against available dmes, and when a matching production is fired, the actions of its action side are executed, changing the set of available dmes. As cognitive primitives, however, they are quite different. ACT-R productions function like goal-driven schemas for accessing information in perceptual-motor buffers and long-term declarative memory (Anderson, 2007). By contrast, 4CAPS productions function like constraints on dmes, activating those that are consistent with each other and suppressing those that are inconsistent with each other (Just and Varma, 2002). As cognitive primitives, 4CAPS make metaphysical claims that are closer to those of the weighted links of IAC networks (Goldman and Varma, 1995). This commensurability arises because at their highest levels, both 4CAPS and IAC networks understand cognition as a form of constraint satisfaction.

To take another example, connectionist architectures include microfeature vectors as basic representations. However, this computational mechanism implements very different cognitive primitives in localist vs. distributed connectionist architectures. Localist representations gain meaning through denotation – each unit codes for one and only one referent (Page, 2000; Bowers, 2009). By contrast, in distributed representations, each unit contributes to the representation of multiple referents, and reference is via similarity (Hinton et al., 1986). The difference is so contentious that some advocates of distributed representations have claimed that localist representations have no place in connectionist architectures at all (Plaut and McClelland, 2010). As cognitive primitives, distributed connectionist representations make metaphysical claims that are closer to those of the episodic traces of exemplar architectures built upon the convolution and correlation operations (Eich, 1985; Murdock, 1993; Plate, 1995).

THE HIGHEST LEVEL

If the lowest level specifies the minutiae of cognitive information processing, it is at the highest level that a cognitive architecture offers its broadest characterization of thinking. This section first reviews Marr's seminal description of this level, which emphasizes its objective meaning. It then articulates the subjective meaning of this level.

FUNCTIONAL SPECIFICATION

In Marr's decomposition, the highest level of a cognitive theory is the "computational theory" it offers. This is a *functional specification* of cognition "as a mapping from one kind of information to another" where "the abstract properties of this mapping are defined precisely" (Marr, 1982, p. 24). The details of how this mapping is implemented are left to lower levels.

Marr argued for the existence of the highest level through a critical review of vision research following World War II. Empirical studies had revealed much about the implementation of the visual system. Emphasizing the lowest level of theoretical description was advocated most strongly in Barlow's (1972) "neural doctrine," which asserted that "a description of the activity of individual nerve cells is a sufficient basis for understanding the function of the visual perception" (p. 380). Marr's review came to a very different conclusion: although neuroscience theories were revealing the *what* and *how* of vision, they were not explaining the *why*.

> Suppose, for example, that one actually found the apocryphal grandmother cell. Would that really tell us anything much at all? It would tell us that it existed – Gross's hand-detectors tell us almost that – but not *why* . . . such a thing may be constructed from the outputs of previously discovered cells.
>
> (Marr, 1982, p. 15)

The limitations of theorizing only at the lowest level are not particular to neuroscience (Anderson, 1972; Brooks, 1991). Marr argued that every cognitive science theory must include a highest level that specifies the function of its domain. He gave one example of a high-level theory from mathematics. The field axioms specify the abstract properties of algebraic expressions, such as the commutativity of addition, but are silent on low-level matters of implementation, such as how numbers are represented (Roman numerals? base-10? base-2?). Marr gave two examples from cognitive science. The first was Gibson's (1979) "ecological" theory of visual perception, which defines the function of visual perception – to enable organisms to navigate their ecological environments – independently of the computational details of how that function is implemented. The second example was Chomsky's (1965) theory of linguistic "competence," which defines the set of language structures. Exactly how these structures are mapped or computed from inputs such as words or sounds – the data structures, parsing algorithms, memory systems, and so on – is left to a lower-level theory of linguistic "performance."

PROCESSING STYLE

Marr's characterization of the highest level as a functional mapping emphasizes its objective meaning. It does not capture its subjective meaning – the broadest ways in which cognitive theories make their domains comprehensible to cognitive scientists. This can be seen by returning to the example of the field axioms. Although they specify the form of algebraic expressions, they do not completely capture the meaning of algebra in the lives of mathematicians. To claim otherwise is to believe that Diophantus, Brahmagupta, and the other great algebraists who lived before their formulation did not understand the subject to which they contributed so much.

The subjective meaning of the highest level is the *processing style* it attributes to cognition. Although missing in Marr's analysis of cognitive theories, it is nascent in Newell's and Anderson's analyses of cognitive architectures, as we will see next. This is perhaps not surprising. Cognitive architectures are computational formalisms – are programing languages. Programing languages cluster into "paradigms" or "families" based on their underlying model of computation. Imperative languages such as C model computation in terms of the von Neumann architecture, functional languages such as Lisp in terms of the lambda calculus, logical languages such as Prolog in terms of logical inference, and so on (Bergin and Gibson, 1996; Wexelblat, 1981). To understand a programing language is to think through its model of computation, and to write programs that express this model rather than fight against it. Similarly, to understand a cognitive architecture at the highest level is to think through its model of computation – its processing style – and to write models that express it in their cognitive information processing.

We next consider two example processing styles. That they are each implemented by multiple cognitive architectures gives evidence of their generality.

Rationality and optimality

A number of cognitive scientists have proposed that cognitive information processing is, at its highest level, *rational*. This is true of Newell's (1982) "knowledge level," with its accompanying "principle of rationality," and Anderson's (1990) "rational level."

Rationality is a processing style with a pedigree: many of the most elegant theories in science appeal to the *optimality* of the natural world. One example is Fermat's principle of least time, which states that "of all the possible paths that it might take to get from one point to another, light takes the path that requires the shortest time" (Feynman et al., 2011, pp. 26-3). This principle can be stated and applied independently of the details *how* the optimal path is computed, which are left to a lower level theory. Optimality principles seem to give a purpose to – explain the *why* of – the natural world. Perhaps for this reason, theories that appeal to optimality are often judged to be of greater esthetic merit, another component of their subjective meaning (McAllister, 1996).

Different cognitive architectures implement the rational processing style using very different lower levels. Soar adopts a procedural notion of rationality, learning from prior problem solving new procedural knowledge to optimize the speed of future of problem solving. ACT-R adopts a Bayesian notion of rationality, learning statistics over prior experiences to take actions that maximize expected utility in the future (Anderson, 2007). That the rational processing style can be implemented by different sets of cognitive primitives demonstrates the quasi-independence of the highest and lowest levels. As Anderson (1990, p. xi) writes, "a rational analysis can stand on its own," independent of the cognitive primitives of "an architectural theory."

Constraint satisfaction

A number of cognitive architectures characterize cognition as a form of *constraint satisfaction*. The next cognitive state is not computed directly, as it is in symbolic architectures that utilize "forward chaining" and connectionist architectures where activation flows in a "feedforward" direction. Rather, a set of constraints defines the landscape of possible cognitive states, an objective function defines the "goodness" of each one, and the next cognitive state is the one that maximizes the objective function subject to the constraints. In "hard" constraint satisfaction, the next cognitive state must satisfy all of the constraints. It is typically implemented by architectures that utilize symbolic computational mechanisms, such as marker-passing networks (Waltz, 1975; Fahlman, 1979) and symbolic programing languages (Sussman and Steele, 1980). In "soft" constraint satisfaction, the next cognitive state satisfies many, but not necessarily all, of the constraints. It is implemented by connectionist networks that employ distributed representations and thermodynamic control structures (e.g., settling, simulated annealing), such as Hopfield (1982) networks and Boltzmann machines (Ackley et al., 1985). It is also implemented by hybrid architectures that utilize both symbolic and connectionist computational mechanisms at their lowest levels, including Pandemonium (Selfridge, 1959), IAC networks (McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982; Kintsch, 1988), classifier systems (Holland et al., 1986), and 4CAPS (Just and Varma, 2007). As these examples demonstrate, the constraint satisfaction processing style is quasi-independent of the cognitive primitives that implement it.

THE MIDDLE LEVEL, BRIEFLY

There is also a middle level to cognitive theories and cognitive architectures. We briefly analyze its objective and subjective meanings here, and direct the interested reader to Varma (2011) for a fuller explication.

Marr defines the middle level objectively, as "the representation for the input and output and the algorithm to be used to transform one into the other" (pp. 24–25). Newell (1989, p. 404) gives a similar definition, colored by his advocacy of symbolic architectures: "the symbol level is that of data structures with symbolic operations on them, being carried out under the guidance of plans, programs, procedures, or methods" (p. 404). Other objective characterizations include Pylyshyn's (1984) "symbolic" level and Anderson's (1990) "algorithm" level. What is common to all is the proposal that at the middle level, the computational mechanisms of the lowest level combine into *data structures* and *algorithms*, to implement the functional specification of the highest level.

The middle level has a parallel subjective meaning. It is where the cognitive primitives of the lowest level combine to process information in an architecture's characteristic style. We call these combinations *idioms* (Lallement and John, 1998; Jones et al., 2007). They help cognitive scientists understand cognition in at least two ways.

First, idioms possess *pragmatic value*. Some problems occur over and over again during model construction. Each problem can be solved by multiple combinations of cognitive primitives. The question, then, is which combination is "best"? Idioms answer this question. They are patterns of cognitive primitives that solve recurring problems in a canonical manner, one consistent with the overall processing style of an architecture (Chase and Simon, 1973; Gamma et al., 1995). For example, when constructing connectionist models of complex cognition (e.g., sentence comprehension), certain problems occur that cannot be solved at the lowest level. One such problem is the representation of variable bindings (e.g., when computing the agreement between two phrases). It is often solved using the CONJUNCTIVE CODING idiom, whereby by a population of units is defined, one for each possible combination of feature values (Hinton et al., 1986; Touretzky and Hinton, 1988). Another such problem is the representation of recursively embedded information (e.g., syntactic structures). This problem cannot be solved at the lowest level because the basic representations, microfeature vectors, are "flat." Connectionist architectures solve this problem using a variety of idioms at the middle level. In feedforward architectures, the TENSOR idiom can be used to encode structured information using vector representations (Smolensky, 1990). In recurrent architectures, the STARTING SMALL idiom – biasing early training toward simpler structures and later training toward complex structures – can be used to learn structured representations within hidden layers (Elman, 1993). This raises the question of why different connectionist architectures solve the recursive embedding problem using different idioms. The reason is that each idiom solves the problem in a manner consistent with its architecture's metaphysical claims at the lowest level and its processing style at the highest level. Although feedforward and recurrent architectures have similar cognitive primitives, they realize different processing styles, and therefore solve the recursive embedding problem using different idioms.

The second contribution that idioms make to the subjective meaning of the middle level is to *enhance communication* between cognitive scientists. They help cognitive scientists understand computational models written by other members of the architectural community. These models are seen not as tangles of cognitive primitives ("spaghetti code"), but rather as patterns signifying the problems that arose during model construction, and how they were solved. Idioms also increase the efficiency of communication. Cognitive scientists who belong to the same architectural community know the same idioms. Therefore, their discussions can utilize the succinct vocabulary of the middle level, and not default to the verbose vocabulary of the lowest level.

IMPLICATIONS

We have articulated the objective meanings of the different levels of cognitive architecture, following analyses originated by Marr, Newell, Pylyshyn, and Anderson. We have also identified the subjective meaning of each level – the understanding it brings cognitive scientists of cognition (see **Table 1** for a summary). Importantly, the objective meaning of a level is quasi-independent of its subjective meaning: one cannot entirely replace the other ("independence"), though they can mutually constrain each other ("quasi").

Here, we draw several implications of this analysis. We first argue that the subjective meanings of different levels of a cognitive architecture are also quasi-independent of one another. We next argue against reducing higher levels to lower levels, for example, in the name of parsimony, because this would lose the subjective meaning unique to higher levels. This would also needlessly limit the flexibility of cognitive scientists to choose the architectural level most relevant for understanding the phenomena of interest to them. We conclude by considering the implications of the multiple meanings and multiple levels of cognitive architecture for understanding the current, disunified state of theoretical cognitive science.

QUASI-INDEPENDENCE

We have seen that the objective meaning of each level is quasi-independent of its subjective meaning. Returning to a previous example, ACT-R and 4CAPS have similar objective meanings at the lowest level, with both including productions as basic operators. However, productions have very different subjective meanings in the two architectures – are very different cognitive primitives. They function as goal-driven schemas for accessing relevant information in ACT-R, whereas they function as constraints between representations in 4CAPS.

Table 1 | Summary of the multiple meanings and multiple levels of cognitive architecture.

Level	Objective meaning	Subjective meaning
Highest	Functional specification: mapping from perceptual-cognitive inputs to cognitive-motor outputs	Processing style: model or paradigm of computation
Middle	Data structures and algorithms: combinations of computational mechanisms that implement the functional specification	Idioms: combinations of cognitive primitives that solve problems that recur during model construction in a manner consistent with the processing style
Lowest	Computational mechanisms: basic representations, basic operators, and control structure of cognitive information processing	Cognitive primitives: specify a metaphysics for cognition

A natural question is the relation between the meanings of different levels. Simon (1996) observed that complex systems tend to be organized hierarchically, with components at higher levels being *nearly decomposable* into components at lower levels. Marr (1982) argued that, for the case of cognitive theories, the objective meanings of different levels are quasi-independent.

> The three levels are coupled, but only loosely. The choice of an algorithm is influenced for example, by what it has to do and by the hardware in which it must run. But there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two.
>
> Marr, 1982 (pp. 24–25)

The subjective meanings of different levels are also quasi-independent. The processing style of the highest level is quasi-independent of the idioms of the middle level, which are quasi-independent of the cognitive primitives of the lowest level. Here "quasi-independence" means that the subjective meanings of different levels can mutually constrain each other ("quasi"), but cannot entirely replace each other ("independence"). We argue for this proposal indirectly, by drawing its implications and providing evidence for them from the history of cognitive architecture.

AGAINST REDUCTION

One implication of the proposal that the subjective meanings of different levels are quasi-independent is that higher levels cannot be entirely reduced to lower levels. This implication is provocative because it flies in the face of parsimony, the standard esthetic criterion in science. This is the preference for simpler theories over more complex theories, all other things being equal (McAllister, 1996). For example, the Ptolemaic and Copernican theories provided comparable accounts of the structure of the solar system – of the observed movements of planets. The Copernican theory came to be preferred in part because it was simpler, i.e., did not require *ad hoc* assumptions about epicycles. This implication is also provocative because it is antithetical to reduction, the standard unification strategy in science (Oppenheim and Putnam, 1958). When higher-level theories are reduced to lower-level theories, macroscopic phenomena come to be explained as emergent properties of microscopic phenomena. An example of a successful reduction is Pauling's explanation of the chemical bond in terms of quantum mechanics, a physical theory. Within cognitive science, this strategy has been advocated most forcefully by "eliminative" reductionists (Churchland, 1981). They argue that higher-level theories are "folk psychological" – approximate at best and incorrect at worst – and should be reduced away to lower-level theories of neural information processing.

There are two reasons why higher levels cannot be entirely reduced to lower levels. The first is that reduction is *underdetermined*. The subjective meanings of different levels are quasi-independent, and in particular the same processing style can be realized by different sets of cognitive primitives that make distinct, even incommensurable metaphysical claims. Therefore, there is no "best" reduction. Returning to a previous example, both ACT-R and Soar implement the rational processing style, but they do so using very different cognitive primitives. To select the next operator to perform, ACT-R uses Bayesian cognitive primitives that maximize expected utility. By contrast, Soar uses set-theoretic primitives, asserting preferences to (partially) order candidate operators and then selecting the most preferred one. Should the rational processing style be reduced to the Bayesian cognitive primitives of ACT-R or the set-theoretic primitives of Soar?

The second reason that reduction fails is because it is *lossy*. In his famous paper "More is Different," Anderson (1972) argued that condensed matter physics cannot be entirely reduced to particle physics because "at each level of complexity entirely new properties appear" (p. 393). Similarly, because the subjective meaning of a higher architectural level is quasi-independent of the subjective meaning of a lower level, some of its unique meaning will be necessarily lost during reduction. Returning to a previous example, the STARTING SMALL idiom solves the problem of representing recursive embeddings for recurrent connectionist architectures. If this idiom is reduced away – replaced everywhere in the literature with its defining combination of cognitive primitives – then its pragmatic value would be lost. Cognitive scientists trying to comprehend the sentence processing model of Elman (1993) would not understand the theoretical claim behind decreasing the proportion of simple structures and increasing the proportion of complex structures over training. They would incorrectly dismiss it as a "hack." The communicative value of the idiom would also be lost. For example, consider connectionists discussing the modeling of problem solving. They would not be able to discuss the representation of plans, which are recursively embedded structures, in terms of the STARTING SMALL idiom. Rather, they would be forced to converse at the lowest level, in the language of cognitive primitives, increasing the ambiguity and verbosity of their communication.

APPROPRIATENESS

Different levels do not just convey different subjective meanings. They also explain cognition at different scales. This provides cognitive scientists with the flexibility to select the most *appropriate* level for understanding their phenomena of interest. Reducing away higher levels in the name of parsimony or unification would needlessly sacrifice this flexibility.

That different theories explain at different scales, and that scientists choose the most appropriate level given the phenomena they seek to understand, is evident in other sciences. For example, Carnot formulated classical thermodynamics to explain macroscopic phenomena such as the operation of heat engines. A half century later, Maxwell, Boltzmann, and Gibbs reduced its laws to those of classical mechanics, applied at the molecular level. Their statistical thermodynamics did not reduce away the older theory; scientists did not stop speaking of "temperature" and start speaking only of "mean molecular kinetic energy." Rather, scientists gained an additional level of explanation, and the flexibility to choose the most appropriate one given the scale of the phenomena to be understood.

Similarly, cognitive scientists select the level most appropriate for understanding the cognitive phenomena at hand. An important factor in this selection is the temporal scale or frequency of the phenomena (Newell and Simon, 1972; Pylyshyn, 1984). Higher levels are more appropriate for understanding cognitions that unfold over longer time scales, such as problem solving, whereas lower levels are more appropriate for understanding cognitions that unfold over shorter time scales, such as word recognition. If the level selected is too high, then the explanation it offers will be too coarse – will be insensitive to the moment-by-moment time course. If the level selected is too low, then the converse problem will arise: cognitive scientists will be forced to make overly detailed claims about moment-by-moment processing that cannot be evaluated against empirical data.

IDENTIFIABILITY

We conclude by considering the implications of the analysis offered here for progress toward "better" cognitive architectures. Many cognitive scientists are committed to a fallibilist approach to scientific progress, where competing theories are put to empirical tests, corroborated theories are retained, and falsified theories are dismissed (Popper, 1963). And yet historically, it has proven difficult to select between competing cognitive theories and cognitive architectures on empirical grounds (Newell, 1973b; Hintzman, 2011). [There are some exceptions. For example, that humans can learn linearly inseparable concepts but perceptrons cannot (Minsky and Papert, 1969) was used to falsify this particular architecture.]

This difficulty is compounded by the *problem of identifiability*. Cognitive architectures are computational formalisms, and most are Turing-equivalent in their computational power. That is, they can express computational models that implement the same functions from perceptual-cognitive inputs to cognitive-motor outputs. Because of their computational equivalence, we cannot select between them based on the "competence" of their computational models. It has been argued that although competing architectures support models that compute the same input–output functions, these models exhibit different "performance" characteristics – different temporal profiles, error distributions, and so on. It might be possible to select the architecture whose models' performance characteristics most closely resemble those of humans, and in this way make progress (Pylyshyn, 1984; Newell, 1990). However, this strategy appears to be undercut by "mimicry" theorems showing that architectures that adopt even diametrically opposed computational mechanisms (i.e., symbolic vs. spatial representations, serial vs. parallel control) can express models that exhibit identical performance characteristics (Townsend, 1974; Anderson, 1978).

One solution to these problems is to abandon the fallibilism of Popper (1963) for the methodology of scientific research programmes proposed by Lakatos (1970). This solution was proposed by Newell (1990) and has been developed in great detail by Cooper (2006, 2007).

The analysis offered here points to an alternative understanding of why progress toward "better" cognitive architectures has been so slow. Comparisons between competing architectures are typically conducted in a particular domain, for example, sentence comprehension, and at a particular level, typically the lowest. Such comparisons are often compromised by the failure to consider appropriateness. If the chosen level is appropriate for modeling the chosen domain in one architecture but not another, then that architecture will be judged as "better." However, if a different level had been chosen, then the choice might have been reversed. More generally, the fallibilist approach cannot ensure progress toward "better" cognitive architectures if appropriateness is ignored.

For example, consider the long-running debate between proponents of symbolic vs. connectionist architectures. Are productions superior to weighted links and activation functions for modeling sentence comprehension, as proponents of symbolic architectures argue? Notice that the phrasing of this comparison is at the lowest level (productions, weighted links, activation functions). This is the appropriate level for addressing the information processing requirements of sentence comprehension – recursive embeddings, variable bindings – in symbolic architectures. However, it is inappropriate for addressing these requirements in connectionist architectures. As we saw above, it is at the middle level that connectionist architectures provide idioms for recursive embeddings (e.g., STARTING SMALL) and variable bindings (e.g., CONJUNCTIVE CODING). And thus it is not surprising that such comparisons have generally been indeterminate. When the ability of connectionist architectures to support models of sentence comprehension is evaluated at the appropriate level, then the result can be much more informative (Steedman, 1999).

More generally, when cognitive scientists use cognitive architectures to understand cognitive phenomena, they select the level most appropriate for the phenomena to be explained. This level is different for different architectures and for different domains. Marr's analysis was seminal in revealing this complexity, and continues to be an important component of the meta-theory of cognitive science.

ACKNOWLEDGMENTS

I thank Marcel Just, Ken Koedinger, Susan Goldman, Mike Byrne, Kyle Cave, Wes Sherman, Jay Konopnicki, and Chad Marsolek for enlightening discussions of these ideas over the years.

REFERENCES

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cogn. Sci.* 9, 147–169. doi: 10.1207/s15516709cog0901_7

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychol. Rev.* 85, 249–277. doi: 10.1037/0033-295X.85.4.249

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195324259.001.0001

Anderson, P. W. (1972). More is different. *Science* 177, 393–396. doi: 10.1126/science.177.4047.393

Barlow, H. (1972). Single units and sensation: a neuron doctrine for perceptual psychology. *Perception* 1, 371–394. doi: 10.1068/p010371

Bergin, T. J., and Gibson, R. G. (1996). *History of Programming Languages-II*. Reading, MA: Addison-Wesley.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychol. Rev.* 116, 220–251. doi: 10.1037/a0014462

Brooks, R. A. (1991). Intelligence without representation. *Artif. Intell.* 47, 139–159. doi: 10.1016/0004-3702(91)90053-M

Campbell, F. W., and Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *J. Physiol.* 197, 551–566.

Chase, W. G., and Simon, H. A. (1973). Perception in chess. *Cogn. Psychol.* 4, 55–81. doi: 10.1016/0010-0285(73)90004-2

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *J. Philos.* 78, 67–90. doi: 10.2307/2025900

Cooper, R. P. (2006). Cognitive architectures as Lakatosian research programs: two case studies. *Philos. Psychol.* 19, 199–220. doi: 10.1080/09515080500462388

Cooper, R. P. (2007). The role of falsification in the development of cognitive architectures: insights from a Lakatosian analysis. *Cogn. Sci.* 31, 509–533. doi: 10.1080/15326900701326592

Crick, F., and Asanuma, C. (1986). "Certain aspects of the anatomy and physiology of the cerebral cortex," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, *Psychological and Biological Models*, eds J. L. McClelland, D. E. Rumelhart, and The PDP Research Group (Cambridge, MA: MIT Press), 333–371.

Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychol. Rev.* 92, 1–38. doi: 10.1037/0033-295X.92.1.1

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Elman, J. L. (1993). Leaning and development in neural networks: the importance of starting small. *Cognition* 48, 71–99. doi: 10.1016/0010-0277(93)90058-4

Fahlman, S. E. (1979). *NETL: A System for Representing and Using Real-World Knowledge*. Cambridge, MA: MIT Press.

Feynman, R. P., Leighton, R. B., and Sands, M. (2011). *The Feynman Lectures on Physics. The New Millennium Edition*, Vol. I, *Mainly Mechanics, Radiation, and Heat*. New York: Basic Books.

Fodor, J. A., and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71. doi: 10.1016/0010-0277(88)90031-5

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns*. Reading, MA: Addison-Wesley.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin.

Goldman, S. R., and Varma, S. (1995). "CAPping the construction–integration model of discourse comprehension," in *Discourse Comprehension: Essays in Honor of Walter Kintsch*, eds C. Weaver, S. Mannes, and C. Fletcher (Hillsdale, NJ: Erlbaum), 337–358.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artif. Intell.* 46, 47–75. doi: 10.1016/0004-3702(90)90004-J

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). "Distributed representations," in *Parallel Distributed Computing: Explorations in the Microstructure of Cognition*, Vol. 1, *Foundations*, eds D. E. Rumelhart, J. L. McClelland, and PDP Research Group (Cambridge, MA: MIT Press), 77–109.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychol. Rev.* 93, 411–428. doi: 10.1037/0033-295X.93.4.411

Hintzman, D. L. (2011). Research strategy in the study of memory: fads, fallacies, and the search for the "coordinates of truth." *Perspect. Psychol. Sci.* 6, 253–271. doi: 10.1177/1745691611406924

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Hopfield, J. (1982). Neuronal networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.

Jones, R. M., Lebiere, C., and Crossman, J. A. (2007). "Comparing modeling idioms in ACT-R and soar," in *Proceedings of the 8th International Conference on Cognitive Modeling*, eds R. L. Lewis, T. A. Polk, and J. E. Laird (Oxford: Taylor & Francis/Psychology Press), 49–54.

Just, M. A., and Varma, S. (2002). A hybrid architecture for working memory. *Psychol. Rev.* 109, 54–64. doi: 10.1037/0033-295X.109.1.55

Just, M. A., and Varma, S. (2007). The organization of thinking: what functional brain imaging reveals about the neuroarchitecture of cognition. *Cogn. Affect. Behav. Neurosci.* 7, 153–191. doi: 10.3758/CABN.7.3.153

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction–integration model. *Psychol. Rev.* 95, 163–182. doi: 10.1037/0033-295X.95.2.163

Laird, J. E. (2012). *The Soar Cognitive Architecture*. Cambridge, MA: MIT Press.

Lakatos, I. (1970). "Falsification and the methodology of scientific research programmes," in *Criticism and the Growth of Knowledge*, eds I. Lakatos and A. Musgrave (Cambridge: Cambridge University Press), 91–196.

Lallement, Y., and John B. E. (1998). "Cognitive architecture and modeling idiom: an examination of three models of the Wickens task," in *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, eds M A. Gernsbacher and S. J. Derry (Hillsdale, NJ: Erlbaum), 597–602.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Presentation of Visual Information*. New York: W. H. Freeman. doi: 10.1016/0042-6989(82)90079-7

McAllister, J. W. (1996). *Beauty and Revolution in Science*. Ithaca, NY: Cornell University Press.

McClelland, J. L., and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: part 1. An account of basic findings. *Psychol. Rev.* 88, 375–407. doi: 10.1037/0033-295X.88.5.375

Minsky, M., and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Murdock, B. B. (1993). TODAM2: a model for the storage and retrieval of item, associative, and serial-order information. *Psychol. Rev.* 100, 183–203. doi: 10.1037/0033-295X.100.2.183

Newell, A. (1973a). "Production systems: models of control structures," in *Visual Information Processing*, ed. W. G. Chase (New York: Academic Press), 463–526.

Newell, A. (1973b). "You can't play 20 questions with nature and win: projective comments on the papers of this symposium," in *Visual Information Processing*, ed. W. G. Chase (New York: Academic Press), 283–308.

Newell, A. (1982). The knowledge level. *Artif. Intell.* 18, 87–127. doi: 10.1016/0004-3702(82)90012-1

Newell, A. (1989). "Putting it all together," in *Complex Information Processing: The Impact of Herbert A. Simon*, eds D. Klahr and K. Kotovsky (Hillsdale, NJ: Erlbaum), 399–440.

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

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *J. Exp. Psychol. Learn. Mem. Cogn.* 10, 104–114. doi: 10.1037/0278-7393.10.1.104

Oppenheim, P., and Putnam, H. (1958). "Unity of science as a working hypothesis," in *Minnesota Studies in the Philosophy of Science*, Vol. II, *Concepts, Theories, and* *the Mind–Body Problem*, eds H. Feigl, M. Scriven, and G. Maxwell (Minneapolis: University of Minnesota Press), 3–36.

Page, M. (2000). Connectionist modelling in psychology: a localist manifesto. *Behav. Brain Sci.* 23, 443–512. doi: 10.1017/S0140525X00003356

Plate, T. A. (1995). Holographic reduced representations. *IEEE Trans. Neural Netw.* 6, 623–641. doi: 10.1109/72.377968

Plaut, D. C., and McClelland, J. L. (2010). Locating object knowledge in the brain: comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychol. Rev.* 117, 284–290. doi: 10.1037/a0017101

Pollack, J. (1990). Recursive distributed representations. *Artif. Intell.* 36, 77–105. doi: 10.1016/0004-3702(90)90005-K

Popper, K. R. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York: Harper and Row.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science.* Cambridge, MA: MIT Press.

Raaijmakers, J. G. W., and Shiffrin, R. M. (1981). Search of associative memory. *Psychol. Rev.* 88, 93–134. doi: 10.1037/0033-295X.88.2.93

Rumelhart, D. E., Hinton, G. E., and McClelland, J. L. (1986). "A general framework for parallel distributed processing," in *Parallel Distributed Computing: Explorations in the Microstructure of Cognition*, Vol. 1, *Foundations*, eds D. E. Rumelhart, J. L. McClelland, and PDP Research Group (Cambridge, MA: MIT Press), 45–76.

Rumelhart, D. E., and McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychol. Rev.* 89, 60–94. doi: 10.1037/0033-295X.89.1.60

Rumelhart, D. E., and McClelland, J. L. (1986). "PDP models and general issues in cognitive science," in *Parallel Distributed Computing: Explorations in the Microstructure of Cognition*, Vol. 1, *Foundations*, eds D. E. Rumelhart, J. L. McClelland, and PDP Research Group (Cambridge, MA: MIT Press), 110–146.

Selfridge, O. G. (1959). "Pandemonium: a paradigm for learning," in *Proceedings of the Symposium on Mechanisation of Thought Processes*, eds D. V. Blake and A. M. Uttley (London: H. M. Stationary Office), 511–529.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science* 237, 1317–1323. doi: 10.1126/science.3629243

Simon, H. A. (1996). *The Sciences of the Artificial*, 3rd Edn. Cambridge, MA: MIT Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artif. Intell.* 46, 159–216. doi: 10.1016/0004-3702(90)90007-M

Steedman, M. (1999). Connectionist sentence processing in perspective. *Cogn. Sci.* 23, 615–634. doi: 10.1207/s15516709cog2304_10

Sussman, G. J., and Steele, G. L. Jr. (1980). Constraints – a language for expressing almost-hierarchical descriptions. *Artif. Intell.* 14, 1–39. doi: 10.1016/0004-3702(80)90032-6

Touretzky, D. S., and Hinton, G. E. (1988). A distributed connectionist production system. *Cogn. Sci.* 12, 423–466. doi: 10.1207/s15516709cog1203_4

Townsend, J. T. (1974). "Issues and models concerning the processing of a finite number of inputs," in *Human Information Processing: Tutorials in Performance and Cognition*, ed. B. H. Kantowitz (Hillsdale, NJ: Erlbaum), 133–168.

Varma, S. (2011). The design and evaluation of cognitive architectures. *Cogn. Sci.* 35, 1329–1351. doi: 10.1111/j.1551-6709.2011.01190.x

Waltz, D. L. (1975). "Understanding scenes with shadows," in *The Psychology of Computer Vision*, ed. P. H. Winston (New York: McGraw-Hill), 19–91.

Weinberg, S. (1993). *Dreams of a Final Theory*. New York: Pantheon Books.

Wexelblat, R. L. (1981). *History of Programming Languages.* New York: Academic Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 February 2014; accepted: 25 April 2014; published online: 19 May 2014.
Citation: Varma S (2014) The subjective meaning of cognitive architecture: a Marrian analysis. Front. Psychol. ***5****:440. doi: 10.3389/fpsyg.2014.00440*
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 22 May 2014
doi: 10.3389/fpsyg.2014.00464

Box-and-arrow explanations need not be more abstract than neuroscientific mechanism descriptions

Edoardo Datteri* **and Federico Laudisa**

Department of Human Sciences, University of Milano-Bicocca, Milano, Italy

Edited by:
Roberto Cordeschi, Sapienza University of Rome, Italy

Reviewed by:
Mitchell Herschbach, California State University Northridge, USA
Gualtiero Piccinini, University of Missouri – St. Louis, USA

Correspondence:
Edoardo Datteri, Department of Human Sciences, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Building U6, Room 4145, 20126 Milano, Italy
e-mail: edoardo.datteri@unimib.it

The nature of the relationship between box-and-arrow (BA) explanations and neuroscientific mechanism descriptions (NMDs) is a key foundational issue for cognitive science. In this article we attempt to identify the nature of the constraints imposed by BA explanations on the formulation of NMDs. On the basis of a case study about motor control, we argue that BA explanations and NMDs both identify regularities that hold in the system, and that these regularities place constraints on the formulation of NMDs from BA analyses, and vice versa. The regularities identified in the two kinds of explanation play a crucial role in reasoning about the relationship between them, and in justifying the use of neuroscientific experimental techniques for the empirical testing of BA analyses of behavior. In addition, we make claims concerning the similarities and differences between BA analyses and NMDs. First, we argue that both types of explanation describe mechanisms. Second, we propose that they differ in terms of the theoretical vocabulary used to denote the entities and properties involved in the mechanism and engaging in regular, mutual interactions. On the contrary, the notion of abstractness, defined as omission of detail, does not help to distinguish BA analyses from NMDs: there is a sense in which BA analyses are more detailed than NMDs. In relation to this, we also focus on the nature of the extra detail included in NMDs and missing from BA analyses, arguing that such detail does not always concern how the system works. Finally, we propose reasons for doubting that BA analyses, unlike NMDs, may be considered "mechanism sketches." We have developed these views by critically analyzing recent claims in the philosophical literature regarding the foundations of cognitive science.

Keywords: functional models, neuroscientific explanation, mechanisms, levels of analysis in neuroscience, regularities in neuroscience

INTRODUCTION

Explanations in the behavioral sciences take on a wide variety of styles. Quite often, especially at the early stages of their discovery, behavioral mechanisms are described without reference to brain areas or neural activity. The system is broken down into a number of interconnected components, each assumed to play an active part in the generation of the behavior to be explained. But no mention is made of what brain area, if any, corresponds to each component. For example, studies on motor control often postulate the existence of a "feedback controller" component in the system that produces motor commands on the basis of trajectory errors, without specifying which part of the target nervous system is presumed to perform this activity. When system components are only characterized on the basis of the activity they perform in the generation of the behavior to be explained, this is often referred to as a "box-and-arrow" (BA from now on) analysis of the system. An example of a BA analysis of motor control is shown in **Figure 1**. However, other behavioral mechanisms are described in terms of anatomically identified brain areas and the associated neural activities. For example, visually guided motor control in humans is thought to involve areas such as the visual cortex, the brain stem, the cerebellum and others. Similarly to BA explanations[1], such neuroscientific mechanism descriptions (from now on NMDs) are often represented in box-and-arrow format in scientific papers (see **Figure 3** for an example); in contrast with what we have termed BA analyses, however, each box stands for a particular brain area or portion of the nervous system. Some brain areas may be linked to an activity performed within the framework of the behavior to be explained (e.g., in navigation when the hippocampus is said to hold a representation of space), but this is not always the case: a part of the nervous system may feature in an NMD even when the precise activity it carries out within the framework of the target behavior is not made explicit.

The theoretical vocabulary used in these two kinds of models is different, at least *prima facie*[2]. NMDs identify components and their organization using the language of neuroscience, which includes terms denoting brain areas, and expressions such as "neural activity," "inhibitory connection," "firing rate" and so

[1] The expressions "BA analysis" and "BA explanation" will be used interchangeably in this article.

[2] The intended meaning of the expression "theoretical vocabulary" will be clarified in Section "The Relationship Between Functional Models and Neuroscientific Mechanism Descriptions."

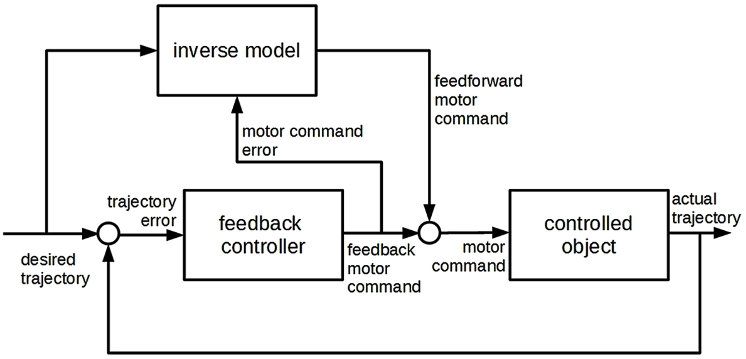


FIGURE 1 | A box-and-arrow model of motor control (adapted from Wolpert et al., 1998).

on. In contrast BA explanations identify components and their organization in terms of a representational or information-processing language. For example, "feedback controllers" are said to produce internal representations of motor commands on the basis of a representation of trajectory error produced by other components. Yet most neuroscientists and philosophers of neuroscience assume that BA analyses and NMDs may be related to, and place constraints on, each other in some cases. Often, explanation of a behavior starts from the formulation of a BA analysis of the target system. Later, this BA analysis is taken as a basis on which to formulate an NMD by seeking out neural components that perform the activities specified in the BA analysis. In other cases, one starts with a NMD featuring interconnected brain areas, going on to identify the specific activities carried out by each: the NMD, in these cases, is taken as a basis on which to formulate a BA model of the system.

Now, it is one thing to assume that some kind of relationship holds between the two types of explanation, but quite another to clarify the nature of this relationship. What kind of constraints do BA analyses place on the formulation of NMDs, and vice versa? This is a question of great importance for neuroscientific research. Understanding the nature of these constraints would provide criteria for deriving NMDs from BA analyses in a principled way, and for testing the latter according to the empirical methods of the neurosciences. It would also contribute significantly to unifying branches of behavioral science, such as cognitive psychology (which typically adopts forms of BA analysis) and basic neuroscience, whose theoretical vocabulary may appear *prima facie* unrelated to each other. The aim of the present article is to take some further steps towards the formulation of such criteria on the basis of a close analysis of a case study on motor control in humans, and by critical examination of views recently expressed in the philosophical literature on the foundations of the cognitive sciences (Piccinini and Craver, 2011; Levy and Bechtel, 2013).

In the selected case study, a BA analysis of motor control was formulated, whose functional structure was claimed to correspond to the structure of a particular mechanism description couched in the vocabulary of neuroscience (Wolpert et al., 1998). The BA explanation and the NMD are described in Sections "On the Structure of Box-and-Arrow Models in Neuroscience" and "On the Structure Of Neuroscientific Mechanism Descriptions," respectively. In Section "The Relationship Between Functional Models and Neuroscientific Mechanism Descriptions" we argue that, in both cases, a number of *regularities* were claimed to hold in the system, although different theoretical vocabulary was used to denote the entities and properties involved in these regularities. We further argue that these regularities played a crucial role in justifying the correspondence between the two explanations. Indeed, the formulation of an NMD proceeded by searching for neural groups whose activities conformed to the relationships expressed in the BA analysis. Thus, the correspondence between the two explanations seemed to consist, in the authors' view, of a correspondence between the regularities expressed in each, while the BA analysis placed constraints on neuroscientific research given that it postulated a number of regularities to be sought out in the neural activity of the system.

In Section "The Relationship Between Functional Models and Neuroscientific Mechanism Descriptions," we also take the selected case study to support a number of claims about the structural similarities and differences between BA analyses and NMDs. As far as the similarities are concerned we argue, consistently with what has been claimed, amongst others, by Piccinini and Craver (2011) that both explanations describe *mechanisms*, given that they refer to system components that interact with each other in a regular fashion[3]. In relation to the structural differences, we examine the claim that BA analyses are *less detailed* (Piccinini and Craver, 2011) or *more abstract* (Levy and Bechtel, 2013) than NMDs. We suggest that BA analyses may provide details that are missing from NMDs, namely, details on the representational roles played by certain neural groups, and that for this reason they may be said to be *richer*, or more detailed than NMDs. For the same reasons we also propose that the notion of *abstractness*, defined as "omission of detail" (Levy and Bechtel, 2013), does not help to define the difference between BA explanations and NMDs. Rather,

[3] For the purposes of the present paper, we provisionally accept the analysis of neuroscientific mechanisms provided by Craver (2007).

the two kinds of explanation differ in relation to the theoretical vocabulary they use to denote system components. By changing theoretical vocabulary, and re-defining BA components in neuroscientific terms, one does not add crucial details about how the mechanism works: one simply describes the same boxes with different vocabulary. A way to add details on how the system works is, rather, to iterate mechanistic analysis on the components of a previously formulated system. This process, often referred to as *decomposition* in the epistemological literature on the cognitive sciences (Bechtel and Richardson, 1993), is to be viewed as distinct from the process of changing the theoretical vocabulary used to describe a mechanism.

We then examine more closely the claim, made by Piccinini and Craver (2011), that BA analyses are elliptical or incomplete versions of neuroscientific mechanism descriptions – *mechanism sketches*, in these authors' terminology – insofar as they leave out crucial details on how the system works. We comment on this view by arguing that the details provided by NMDs and lacking in BA explanations need not concern how the system works. And we also suggest more general reasons for doubting that BA analyses may be considered elliptical versions of NMDs.

Let us begin this discussion by outlining the structure of the BA analysis of motor control proposed by Wolpert et al. (1998).

ON THE STRUCTURE OF BOX-AND-ARROW MODELS IN NEUROSCIENCE

The scientific question addressed by Wolpert et al. (1998) is to understand how human beings control their movements along a desired trajectory – for example, how they successfully move a hand towards a specific object, or move an eye to track a portion of the visual environment. The idea proposed by the authors, and expressed in the BA analysis shown in **Figure 1**, is as follows.

A representation of the desired trajectory is available to the system. Then two components, the "feedback controller" and the "inverse model," produce two motor commands – termed feedback and feedforward motor commands, respectively – that are combined before being sent, as a final motor command, to the "controlled object" (e.g., arm muscles) for execution. Both components produce motor commands, yet they there is a key difference between them. The "feedback controller" produces a motor command on the basis of "trajectory error," i.e., on the basis of the difference between (1) the representation of the desired trajectory, and (2) the representation of the "actual trajectory" followed by the controlled object. This is the classical cybernetic negative-feedback principle (Rosenblueth et al., 1943), which is applied in many self-regulation devices (e.g., in thermostats). A major issue with such feedback-based control loops is the time required to receive feedback information on the actual trajectory. Sensory pathways are very delayed in humans, and a control mechanism based purely on feedback would make the system move too slowly or make too many errors. This was the main reason leading the authors to postulate a sort of short-cut, represented by the "inverse model" module. The function of this module is to generate motor commands on the basis of a representation of the desired trajectory only, with no sensory information available (intuitively, we use an inverse model when moving in our house in the dark). *Feedforward* motor commands are generated much more rapidly than the feedback ones, because they do not need to wait for the arrival and processing of sensory information. When a feedback command is available, it is combined with the feedforward command as earlier stated; otherwise, the system executes the feedforward command only, enabling itself to follow the desired trajectory within a reasonable time-frame.

Clearly, the "inverse model" must be trained before being able to generate the appropriate motor commands. The training signal consists of the representation of motor command error, generated on the basis of trajectory error (we correct our internal model of the house whenever we bump into a wall or piece of furniture that we erroneously believed to be farther away from us).

As mentioned in the Introduction, the authors of this study also formulated a neuroscientific mechanism description, discussed in detail in the next section, which was deliberately made to correspond with the structure described so far. As a basis for understanding the relationship between the two types of analysis, it is worth discussing some aspects of the BA explanation as described here. This explanation implies that there is something in the system which can *represent* desired trajectories. The system can also represent *feedback* and *feedforward motor commands*. Indeed, as explained before, the final motor command is a combination of feedback and feedforward motor commands: a plausible interpretation of this claim would be that the system has internal representations of the two commands, which are then combined into a third representation (the final motor command) driving the effector organs. In addition, the BA analysis refers to a number of functional components, including the "inverse model" and the "feedback controller," presumed to be involved in motor control. These components are parts of the target system that are assumed to fulfill distinct functions within motor control.

This raises the question of what differentiates each component from the others. What is an "inverse model?" The authors of the study suggested that an inverse model is a component that transforms the desired movement trajectory of the controlled object into the motor commands required to attain this movement goal. That is to say that, by claiming that the system has an "inverse model," the authors claimed that there is something in the system that establishes a *regular relationship* between desired trajectories and feedforward motor commands. This regular relationship was not precisely defined in their theory, apart from the claim that each desired trajectory is mapped onto the motor command *that would make the system follow that trajectory*[4]. Similarly, in suggesting that the system has a "feedback controller," the authors claimed that there is something in the system that establishes a regularity between trajectory errors (which, in turn, depend on the difference between desired and actual trajectories) and feedback motor commands: feedback controllers produce motor

[4] More precisely, the regularity associated with the "inverse model" module in the BA diagram links *three* representations to one other, as shown in **Figure 1**: the feedforward motor command depends on the desired trajectory and on the motor command error, which intermittently trains the motor apparatus' internal model.

commands that have the effect of reducing trajectory error[5]. These definitions are rather vague, but they nevertheless impose restrictions on the set of possible regularities associated with the "inverse model" and "feedback controller" components. The other functional components are associated with other regularities. To sum up, this BA model suggests that the system possesses a number of *internal representations* among which certain *regularities* hold.

Note that the BA analysis makes no claim about how each component establishes the corresponding regularity. As often noted in the philosophy of cognitive science, box-and-arrow analysis may be iterated to obtain finer-grained, more detailed BA analyses of the same behavior. See for example **Figure 2**, in which a purely notional BA subanalysis of the "feedback controller" component (not included in Wolpert et al., 1998) is shown. Three components are added, each of which establishes a regularity among additional intermediate representations. The relationship between the BA analysis described above, which we may call M for short, and the richer analysis M′, in which one or more functional components of M are further analyzed and broken down into a box-and-arrow structure, is often defined through appeal to the notion of *decomposition* (Rosenblueth and Wiener, 1945; Cummins, 1985; Bechtel and Richardson, 1993). Clearly, M′ may be further decomposed via an even finer-grained analysis; this process leads to the formulation of a *decomposition hierarchy* of BA explanations.

It is worth stressing here two aspects of BA decomposition that we come back to in the ensuing discussion. First, by decomposing a BA analysis, one obtains a richer BA analysis of the same system, in which further details are added on how the system is thought to work. For example, M is silent on a particular aspect of the functioning of the system, namely on how the "feedback controller" works, simply stating that the "feedback controller" component establishes a regular relationship between two representations. M′ adds details on how this component works, thus adding information on the functioning of the target system. Second, decomposition does not imply a change in the theoretical vocabulary used to describe the organization of the system – for example, specifically with regard to living systems, it does not imply a shift to the vocabulary of neuroscience – or vice versa. This is particularly evident in BA explanations formulated in computer science, in which components establishing regularities between system representations (typically expressed as functions in a given programming language) are analyzed into additional components (sub-functions) that establish regularities connecting additional system representations. Similarly, BA components in the study of cognition are often analyzed (decomposed) by postulating cascades of transformations among intermediate representations. Such a decomposition process does not lead to a change in vocabulary: it simply leads to another, richer, BA explanation. The process of decomposing a BA explanation must be kept conceptually distinct from the process of shifting to another theoretical vocabulary. Later in the paper, we focus on this distinction, arguing that the "translation" of a BA explanation into a neuroscientific MD does not necessarily lead to the addition of crucial details on the working of the system.

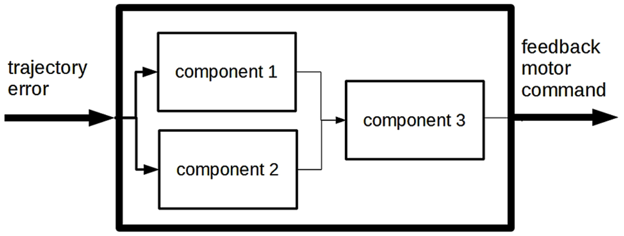


FIGURE 2 | A model of the "feedback controller" component of the BA model.

Furthermore, this distinction enables us to separate two methodological issues, both related to the more general problem of understanding the relationships between functional and mechanistic models in neuroscience. One of these issues is how to characterize the decomposition relationship, that is to say, the relationship holding between an explanation M and another explanation M′ obtained by decomposing from M and expressed using the same theoretical vocabulary. In other words, the issue of identifying the criteria used by scientists to transform previous explanations of a system into richer ones adding crucial details on the working of the system. A different issue is that of characterizing the relationship holding between two explanations of the same behavior formulated using different theoretical vocabularies. The case study analyzed here, as discussed in the next section, provides insights that help to address both questions, although this article is more strongly focused on the second issue.

ON THE STRUCTURE OF NEUROSCIENTIFIC MECHANISM DESCRIPTIONS

Figure 3 shows a diagram formulated to explain how we[6] control our eye movements to track moving portions of a visual scene (the so-called ocular following response or OFR). This motor control function cannot be fulfilled by feedback control only, given that visual feedback in humans is too delayed to enable efficient control of eye movements. The combination of feedback and inverse control, according to the principle described in the previous section, is a more promising approach to explaining this ability. A neuroscientific explanation

[5] The "trajectory error" and the "feedback motor command" representations may well be regarded as the *input* and the *output* of the "feedback controller" component, respectively. However, the use of these terms, although consistent with the present analysis, would require an additional account of what makes something an "input" or an "output" of a BA component – an account which may enable one to understand, e.g., why the "trajectory error" is more properly regarded as an input rather than as an output of the "feedback controller" component. Such an account is not really required for the purposes of the present article. For this reason we provisionally avoid the use of the terms "input" and "output" and only claim that, according to the BA analysis, the "feedback controller" is a component that establishes a regular relationship between the two representations.

[6] This study actually examined monkey OFR, and the neuroscientific mechanism description discussed below refers to areas of the monkey brain. However, the authors suggested that this mechanism description could also shed light on the OFR mechanism in humans. Discussion of the reasons for such a generalization is beyond the scope of the present article.

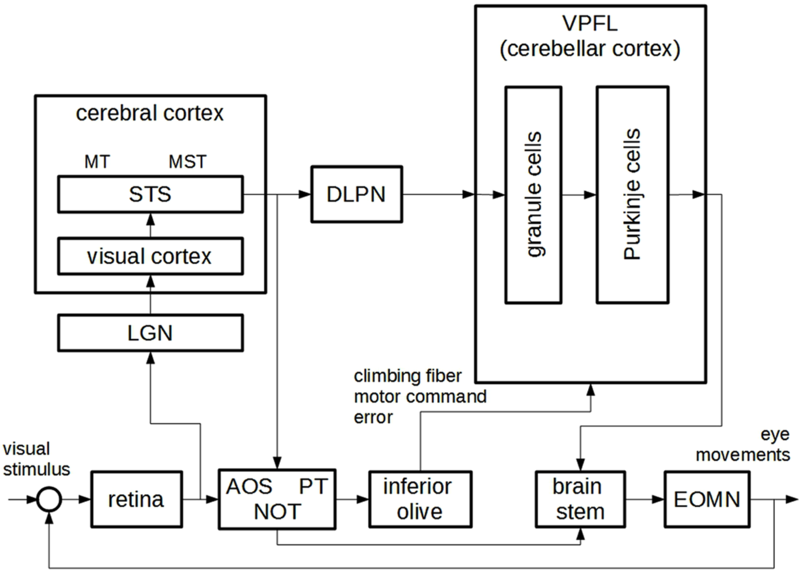


FIGURE 3 | A neuroscientific mechanism description for ocular-following reflex behavior, adapted from (Wolpert et al., 1998). AOS: accessory optic system; PT: pretectum; NOT: nucleus of optic tract; EOMN: extra-ocular motor neurons; LGN: lateral geniculate nucleus; STS: superior temporal sulcus; MT: middle temporal area; MST: medial superior temporal area; DLPN: dorsolateral pontine nucleus; VPFL: ventral paraflocculus.

of OFR was formulated by Wolpert et al. (1998) so as to correspond to the BA diagram represented in **Figure 1**. In particular, the authors claimed that the "inverse model" component corresponded to the component labeled as the ventral paraflocculus of the cerebellar cortex (VPFL) and to a number of additional components, as discussed below. The "feedback controller" corresponded to a set of components including the retina, the lateral geniculate nucleus (LGN), and portions of the visual cortex. Other components mentioned in their mechanistic analysis, as later discussed, cannot be easily mapped onto the BA explanation outlined in the previous section.

A question of primary importance is how the authors justified the claim that the two explanations corresponded to each other, albeit partially. Let us address this question by focusing on the "inverse model" component. The authors claimed that "the VPFL is the major site of the inverse dynamics model of the eye for OFR" (p. 341). This amounted to claiming that the VPFL is responsible for the fact that the activity of a neural group, which we call B for the moment, depends on the activity of another neural group A in the following regular fashion: the activity of B drives eye movements along the trajectory represented in A. In other words, by claiming that the VPFL served as an inverse model for the eye, Wolpert et al. (1998) were suggesting that the VPFL was responsible for a regular connection *of the "inverse model" type* between the activity of two neural groups. Should no neural activity in the brain be dependent in this way on a neural representation of desired trajectories, no inverse model would be claimed to be in the brain. The authors justified the claim that "the VPFL is the major site of the inverse dynamics model of the eye for OFR" by providing neuroscientific evidence for such a regularity. Note that an engineer would argue in the same way that an electromechanical electrical circuit included an inverse model, that is to say, by showing that the electrical activity at some point of a circuit depended on the electrical activity at another point of the circuit in accordance with "inverse model" regularity. In both cases, the fact that the BA analysis specifies *regularities* holding between parts of the system is crucial to understanding the nature of the relationship between BA and neuroscientific (or electromechanical) explanations of the same behavior.

The authors presented some interesting, albeit far from decisive, empirical support for their claim. First of all, the activity of certain VPFL cells – the Purkinje cells, often considered the output of the cerebellum – has been found to be regularly connected with eye movements. In particular, it is known that the activity of the Purkinje cells displays two types of spike, namely simple and complex spikes (SS and CS from now on; see Kandel et al., 2000). The SS are single action potentials. They occur at a relatively high frequency and have been found to correlate with certain aspects of eye movement. According to Wolpert et al. (1998), they may drive eye movements without waiting for the low-frequency arrival of sensory information, thus serving as feedfoward motor commands. Motor correlation has not been found in other neurons projecting from vision-related areas to the cerebellum, namely in the neurons of the dorsolateral pontine nucleus (DLPN) and the medial superior temporal (MST) area. In the authors' view, given the absence of motor correlation and the connections with visual areas, these cells may "provide the desired trajectory information" to the cerebellum. Let us turn now to cerebellar CS. These are large-amplitude spikes followed by bursts of smaller action potentials. In addition they occur at a very low frequency (about 1–3 per second) and, similarly to the SS, display high correlation with eye movements. The authors suggested that the occurrence of a

CS may signal the moment in which a motor command derived from feedback analysis (thus highly delayed, consistently with the low frequency of CS) interferes with the activity of the Purkinje cells and trains the inverse model. This relationship is shown in the BA diagram by the arrow connecting feedback motor commands with the inverse model. These empirical findings were taken by the authors as a basis for conjecturing that "the VPFL is the major site of the inverse dynamics model of the eye for OFR" (p. 341).

At the time of publication of Wolpert et al. (1998), the formulation of an NMD for OFR behavior was still at the early stages of development. However, on the basis of Wolpert et al.'s (1998) report, it is reasonable to believe that they were trying to identify a "neuroscientific version" of the regularities postulated by the BA analysis in the neural activity of the system. These regularities played a crucial role in justifying the correspondence between the BA analysis and the NMD: the VPFL, for example, was conjectured to be the neural site of the "inverse model" component as it supposedly establishes a regularity of the inverse-model type among entities and properties denoted with neuroscientific vocabulary, that is to say, between the neural activity of two neural groups. The justification was sought in the fact that the same regularities, expressed in different theoretical vocabularies, are found in the system. If no reference were made to the regularities postulated by the BA explanation – or if these regularities were specified in a qualitative and imprecise way, or BA components were only described in terms of textual expressions such as "feedforward controller" – it would not be clear how to relate the BA explanation to a neuroscientific mechanism description of the same system.

Note that, according to the regulative principle proposed here, a neuroscientific mechanism description may be formulated from a BA analysis by mapping each functional regularity onto a neural regularity, without adding components. This is the case of neuroscientific mechanism descriptions which reproduce the boxes and arrows of a BA analysis of the same system, while *adding* an indication of the neural structure subserving each functional role. In many cases, however, the shift from a BA explanation to a neuroscientific one is accompanied by a proliferation of neural structures.

This is also the case of the mechanism description analyzed here. Some neural structures internal to the VPFL are represented in the diagram. And various areas, such as the inferior olive and the previously mentioned AOS, PT, and NOT, are not easily mapped onto the BA explanation. Such a proliferation can result from a *decomposition* process, analogous to the process described in the previous section. We have pointed out that BA components may be analyzed into other BA sub-components, organized so as to produce the corresponding regularity (see **Figure 2**). By decomposing explanation M one obtains a richer explanation M′, which includes additional regularities internal to each component expressed in the same theoretical vocabulary. Similarly, the components of a neuroscientific mechanism may be analyzed into subcomponents, expressed using the theoretical vocabulary of neuroscience, and organized so as to produce the corresponding regularity. For example, when analyzing the internal VPFL cerebellar circuitry, additional regularities defining sub-components of the cerebellum may be identified: the richer neuroscientific mechanism is obtained by decomposition of the initial model. A case of decomposition in the study described here concerns the inferior olive, the AOS, the PT and the NOT, which the authors believed to be crucially involved in transforming the representation of the desired trajectory from sensory to motor coordinates, thus contributing to the inverse model transformation. Decomposition adds detail on the inner working of model components – thus, it provides extra detail about how the system is supposed to work – and for this reason it often marks an advance in the study of the modeled system[7]. However, it is worth stressing that it is one thing to "translate" a BA analysis into a neuroscientific mechanism description, and another to decompose the latter in order to obtain a more detailed model; and that – as often acknowledged in the philosophical literature – both BA explanations and NMDs may be decomposed, although they are expressed using different theoretical vocabularies.

THE RELATIONSHIP BETWEEN FUNCTIONAL MODELS AND NEUROSCIENTIFIC MECHANISM DESCRIPTIONS

Let us sum up the claims made so far. In the case study analyzed here, a BA explanation and a NMD were introduced, each outlining a number of components supposedly involved in motor control, and describing their regular interactions. Notably, the neuroscientific mechanism description was claimed to "correspond" to the BA explanation. How did the authors justify this claim? As suggested in the previous sections, a key role in providing such a justification was played by the *regularities* expressed by the two explanations[8]: the authors seemed to follow the regulative principle according to which a BA analysis and a neuroscientific mechanism description "correspond" to each other to the extent that they display the *same* regularities, even though they are expressed in terms of different theoretical vocabularies. Indeed, the authors' search for the neural structures corresponding to each BA component consisted of

[7] As discussed in Footnote 14, this is not to say that decomposition leads always to better explanations. We wish to remain entirely neutral in relation to what level of explanation, or which theoretical vocabulary, is more suited to the purpose of explaining adaptive behavior or answering specific why-questions about it. Our point solely concerns the structural relationship between BA explanations and NMDs.

[8] Whether it is possible to have a theory of what we mean by a "regularity" in the field of neurosciences is, to a large extent, an open question. Woodward's notion of "invariance under interventions" (Woodward, 2003, 2010) is a useful starting point to address this issue. Many authors – including Craver (2007) – have claimed that neuroscientific generalizations are *fragile* and exception-ridden, given that they are subject to a formidable number of boundary conditions. This claim may be taken to pose a serious problem for present analysis: if no robust (as opposed to fragile) generalization may be found in the neural structure of the system, how can the correspondence between the BA explanation and the NMD be justified according to the regulative principle proposed here? This is a legitimate question, needing to be addressed by further analysis; nevertheless, we believe that the regulative principle put forward here is reasonable. Indeed, in our opinion, it is a matter of fact that Wolpert et al. (1998) identified regularities in the neural system. These regularities were claimed to be robust enough to be found in different individuals and at different times. Clarifying how neuroscientific generalizations may be taken as sufficiently robust to license prediction and explanation, despite being subject to a multitude of boundary conditions, is in our opinion an important aim for the epistemological analysis of neuroscience; see the discussion in Datteri and Laudisa (2012).

a search for the regularities characterizing the component, as outlined in the BA analysis, in anatomically connected regions of the brain. What makes something an "inverse model" is the fact that it establishes a particular regularity internal to the system, and what makes something a neural structure serving as an "inverse model" is the fact that it establishes a regularity of the inverse-model type among the activity of different neural groups[9].

This provides a tentative answer to the key issue addressed in this article: what kind of constraints do BA analyses place on the formulation of NMDs, and vice versa? The BA analysis imposes constraints on the formulation of the NMD by postulating a number of regularities to be sought for in the neural activities of the system. Vice versa, the NMD constrains the space of the possible BA analyses of the system by postulating a number of neural regularities. Suppose that the study of a particular aspect of motor control in animals starts from the formulation of an NMD – possibly via the detection of correlations among the firing of different neural groups. Suppose, in addition, that one of these correlations takes a "feedback controller" form – e.g., the firing of neural group B drives muscles so as to reduce firing of neural group A. In that situation it would be reasonable to suppose that the system has a representation of the motor error (the firing of group A) and is able to produce an appropriate motor command to reduce motor error – in simpler terms, that it has a feedback controller. Should the regularity be different (for example, should the firing of A increase over time instead of tending to 0), one would not suppose that the system had a negative feedback controller (it might be thought to have a positive feedback controller instead).

The selected case study also provides a useful basis for assessing the structural similarities and differences between BA analyses and NMDs. As far as the similarities are concerned, both explanations specify a set of regularities supposedly holding in the system, though expressed using different theoretical vocabularies. And for this reason, consistently with Piccinini and Craver (2011), they both describe *mechanisms*. Indeed, if we are willing to consider the structure represented in **Figure 3** as a mechanism description, why not view BA analyses in the same way? Both types of explanation list a number of components suggested to be responsible for the behavior to be explained, and – more crucially – both specify the regular interactions holding among system components via a number of generalizations. The main difference between the two lies in the theoretical vocabulary used, but it is not clear why the choice of a particular theoretical vocabulary should determine whether or not to define something as a "mechanism description."

As already stated, one of the major differences between the two mechanisms concerns the theoretical vocabulary used. The expression "theoretical vocabulary" is used here to denote a set of terms used in a particular discipline, or in a particular area of research, to express scientific theories. Statements regarding the neural activity of particular areas of the nervous system, and the anatomical connections among brain areas, are couched in the theoretical vocabulary of the neurosciences (which includes terms such as "neuron," "neural activity," "cerebellum," "brain," and so on). These terms are not used in what we refer to here as BA explanations[10]. As often pointed out in the philosophical literature on cognitive science, the theoretical vocabulary of BA explanations distinctively includes the term "representation." Indeed, many BA explanations – including the explanation considered here – assume that the target system has a number of representations. And the various functional components, as in the case discussed here, are typically defined by appeal to these representations. Saying that the system has a "feedback controller" component is to make a rather amorphous claim, unless that component is defined more precisely as a component establishing a regular relationship between different representations held by the system. The notion of representation plays a key role in defining the components of a BA analysis and, therefore, in defining a BA explanation.

Do BA analyses and NMDs (also) differ in that the former are *less detailed* or *more abstract* than the latter? Such a position has been taken, amongst others, by Piccinini and Craver (2011), who propose that "functional and mechanistic explanations are not distinct and autonomous from one another precisely because functional analysis, properly constrained, is a kind of mechanistic explanation – an *elliptical* mechanistic explanation" (284). These authors call such elliptical mechanistic explanations *mechanism sketches*; therefore, in their view, functional explanations are mechanism sketches. BA analysis is a type of functional analysis, they propose, because it identifies components on the basis of the functional role they play in the framework of the behavior to be explained[11]. Piccinini and Craver's (2011) identification of BA

[9] Anatomical considerations guide the authors in selecting the regularities which may play a part in the explanation. Indeed, only the regularities holding among anatomically connected parts of the system are typically included in the description of an explanatory mechanism. This is consistent with the view proposed here: to claim that neuroscientific mechanism descriptions formulate regularities holding in the system does not mean to claim that *any* regularity may be included in a description of a mechanism.

[10] According to the definition proposed in Section "Introduction," BA explanations do not specify which parts of the target nervous system are presumed to perform the activities mentioned there. For this reason NMDs are not included in the class of BA explanations, even though (as noted before) NMDs are often represented in a box-and-arrow format in scientific papers. BA and NMDs, as defined here, differ in relation to the theoretical vocabulary used to denote system components. This is not to say that the use of representational terms is incompatible with the use of neuroscientific terms in the same explanation. Indeed, in the course of scientific discovery, researchers often formulate "mixed" explanations using *both* representational and neuroscientific terms. An example can be found in our case study (Wolpert et al., 1998). After describing the BA explanation, the authors propose an analysis of motor control (**Figure 1B** at p. 339) which provides information on the neural localization of the "inverse model" component *only*. This analysis, presented by the authors as an intermediate stage in the formulation of an NMD from the initial BA explanation, uses representational and neuroscientific terms. Nothing, in the view proposed here, rules out the possibility of formulating mixed models of that kind. Rather, the analysis proposed here – focused on the relationship holding between non-mixed BA explanations and NMDs – may also contribute to understanding how mixed analyses are formulated from explanations of the former or the latter kind. Indeed, there are reasons to believe that the mixed analysis in (Wolpert et al., 1998) is obtained from the BA explanation by applying the criteria discussed here to one component only, i.e., by searching for an "inverse-model" type regularity in the neural structure of the target system, and provisionally ignoring the other BA components.

[11] We consider Piccinini and Craver's (2011) views to be relevant to the main points made here, given that we believe that they would classify the analysis represented in **Figure 1** as a functional analysis. Indeed, each of its components is labeled

analyses with mechanism sketches is consistent with their broader view that BA analyses impose constraints on the formulation of neuroscientific mechanism descriptions. We agree with this hypothesis, but not with the hypothesis that BA analyses are elliptical or incomplete mechanism descriptions, for the following reasons.

As already discussed, both functional and neuroscientific mechanism descriptions specify a number of regularities occurring in the system. They differ in terms of the vocabulary used to denote the entities and properties involved in these regularities. Does a change in theoretical vocabulary entail a gain in completeness? One might answer in the affirmative, in light of the fact that the BA analysis does not convey information regarding the brain areas and neural groups underlying the various representations held by the system. The BA explanation, for example, does not specify what neural groups fulfill the role of representing desired trajectories or feedforward motor commands. This may lead one to believe that the BA explanation is less detailed than the NMD. However, it is also true that BA explanations convey information which are absent in NMDs. Indeed, in principle, NMDs need not provide information on the representational functions of the neural groups involved in the mechanism. They may simply identify neural components and define their regular interaction, without claiming, for example, that the firing activity of neural group A is responsible for, encodes, or serves as, a representation of something. Functional information about the system's representational abilities is explicitly provided by BA explanations but may be missing from NMDs[12]. Therefore, BA explanations may convey details that are lacking in NMDs. For these reasons, one may legitimately view NMDs as "more detailed" than BA explanations only by appropriately restricting the term "detail" to refer to "*neural* detail." But the awarding of such an epistemic privilege to neural details requires justification.

These considerations may also be applied to Levy and Bechtel's (2013) views on *abstractness*, defined as "omission of detail." BA explanations omit details provided by NMDs, and vice versa. For this reason, they cannot be ordered on a scale of abstractness without being explicit about the nature of the details at stake (BA explanations are more abstract than NMDs as far as neural details are concerned, and NMDs are more abstract than BA explanations as far as representational details are concerned)[13]. We claim that a better way to define the difference between the two kinds of explanation is to say that they each convey different information (with each abstracting with respect to details of a particular kind) about the target system, by using different theoretical vocabularies.

Let us further elaborate on the nature of the details omitted from BA explanations and provided by NMDs, by recalling that Piccinini and Craver (2011) describe BA analyses as mechanism sketches, which they discuss in the following terms.

> Descriptions of mechanisms [...] can be more or less complete. Incomplete models – with gaps, question-marks, filler-terms, or hand-waving boxes and arrows – are mechanism sketches. Mechanism sketches are incomplete because they leave out crucial details about how the mechanism works. Sometimes a sketch provides just the right amount of explanatory information for a given context (classroom, courtroom, lab meeting, etc.). Furthermore, sketches are often useful guides to the future development of a mechanistic explanation. Yet there remains a sense in which mechanism sketches are incomplete or elliptical (p. 292).

Now, it is one thing to claim that BA explanations are elliptical with respect to neuroscientific mechanism descriptions given that they do not provide information on the neural areas subserving the various representational roles, but another to claim that they are elliptical because they "leave out crucial details about how the mechanism works." By changing theoretical vocabulary, and expressing similar regularities in the language of neuroscience, one does not add crucial details about how the mechanism works: one simply describes the same boxes with a different vocabulary. Answers to questions such as "how does system A work?" take the form of mechanism descriptions; further detail on how system A works is added by decomposing the mechanism description, as illustrated in the previous sections, and not by expressing it with a different vocabulary[14]. If one knows that the target system has a component X, simply defining that component using a

using an expression that denotes a functional role, i.e., "feedback controller" and "inverse model." And the notion of representation, used to characterize the various components, is functional by definition.

[12] This is not to say that NMDs are not *functional*. Indeed many authors, including Piccinini and Craver (2011), have convincingly shown that NMDs are functional according to various interpretations of the term. Not least because they identify neural components that are assumed to play a crucial functional role in the framework of the behavior to be explained. But NMDs do not always convey the particular kind of functional information provided by BA explanations. Specifically, they may include reference to brain areas thought to play a crucial functional role in the mechanism (e.g., the cerebellar cortex) without defining their functional role (e.g., inverse control) or refer to the activity of a particular neural group without providing information on what this activity is thought to represent. As a matter of fact, neuroscientists often *add* this kind of functional detail to purely anatomical NMDs for a range of explanatory or experimental purposes.

[13] Levy and Bechtel's (2013) study concerns the relationship between NMDs and models that "abstract from the structural specifics of a mechanism and represent it in a skeletal, coarse-grained manner." In these models, "the pattern of causal relations within a system is highlighted, while structural aspects of components are suppressed" (241). Such models may be sensibly viewed as more abstract than NMDs, because they are obtained from NMDs by omitting certain sorts of details. However, we doubt that many BA explanations in the cognitive sciences, including the analysis represented in **Figure 1**, may be classified as abstract models of that kind. The reason is that these explanations are obtained from NMDs by omitting neural details *and* by adding representational details: they do something more than representing the pattern of causal relations within the target system. Indeed, it is possible to formulate an abstract model both of the NMD shown in **Figure 3** *and* of the BA analysis shown in **Figure 1**. Suppose, for example, that a "feedback controller" is defined as a component that generates a motor command whose intensity, represented by a real number, is inversely proportional to the intensity of an error, represented by another number, as determined by coefficient *a*. Thus the abstract model of this BA component will take the form $y = ax$. This abstract model is not the same thing as the BA analysis on which it is modeled.

[14] Adding details on how the system works is not the same as providing a better explanation of the target behavior: the added details could well be irrelevant for the given explanatory purpose, for reasons not explored here. We wish to maintain a neutral position with regard to what decomposition level is the most suitable for explaining behavior, and with regard to whether only NMDs can explain. Our sole interest is the relationship between BA analyses and NMDs. For this reason, we do not comment on the claims made by Levy and Bechtel (2013) and by Piccinini and Craver (2011) in relation to the explanatory power of BA analyses and NMDs.

different vocabulary does not *enrich* the mechanism description of the system – rather it *translates* it into another description. This is not to deny that such a translation may mark important progress in the study of the target system, possibly because it paves the way for the application of additional experimental techniques. And it is true that sometimes – as in the present case study – the shift from a functional to a neuroscientific mechanism description is accompanied by a decomposition process. But this need not be always the case. The point to be emphasized is that changing theoretical vocabulary in the description of a system should be clearly distinguished from the process of decomposing a model of the system. This distinction, as already pointed out, has the effect of splitting the question of the relationship between different models of a system into two questions: the first concerns the relationship between models expressed using different vocabularies, while the second concerns the relationship between different levels of the decomposition hierarchy.

This view has another implication in relation to Piccinini and Craver's (2011) claim that BA analyses are mechanism sketches, that is to say, incomplete or elliptical. Are NMDs mechanism sketches too? If so, the notion of mechanism sketch would not help to draw a distinction between BA analyses and NMDs, contrary to the main point made by Piccinini and Craver (2011). It follows that NMDs, for Piccinini and Craver, are not mechanism sketches. And the assumption that BA explanations, as mechanism sketches, are elliptical and incomplete, leads us to conclude that NMDs are not elliptical nor incomplete, namely, that they are *complete* descriptions of a mechanism. It is important to be careful and explicit about the sense in which NMDs may be defined as such, in order to avoid the strong implication that NMDs say *everything* – being complete descriptions – that can be said about a mechanism (e.g., to avoid the implication that the mechanism description relating to long-term potentiation used by Craver, 2002 to illustrate the notion of "mechanism description" says everything about the target mechanism). Here we have claimed that NMDs may omit information about the representational roles of the neural structures in the target system and, that, for this reason, they may sensibly be view as incomplete with respect to BA explanations.

CONCLUSION

The nature of the relationship between box-and-arrow explanations, which do not invoke neural mechanisms, and neuroscientific mechanism descriptions, is a key foundational issue for cognitive science. On the one hand, the opportunity to disregard neural details in the explanation of behavior has in the past been a source of insight and creativity, yielding hypotheses that led to a better understanding of numerous behavioral and cognitive phenomena. On the other hand, the strong increase in detail of analysis, both theoretically and experimentally, on the part of the neurosciences has led to a corresponding increase in the production of models whose cognitive significance, however, is still far from clear and unequivocal. In the present article we have attempted to tackle the question from a foundational viewpoint, by focusing on the nature of the relationship between box-and-arrow, non-neural explanations of behavior, and neuroscientific mechanism descriptions. On the basis of a case study concerning motor control, we first argued that the regularities formulated in box-and-arrow explanations and neuroscientific mechanism descriptions play a crucial role in justifying any "correspondence" between the two. The regularities formulated in BA explanations place constraints on the formulation of NMDs, and vice versa. Then, we made some general claims regarding the similarities and differences between BA analyses and NMDs. As far as the similarities are concerned, consistently with other positions expressed in the literature, we argued that both kinds of explanations describe mechanisms. As far as the differences are concerned, we suggested that the two kinds of explanation differ in terms of the theoretical vocabulary used to denote the entities and properties involved in the mechanism and engaging in regular, mutual interaction. On the basis of the selected case study we also argued, first, that the notion of abstractness, defined as omission of detail, does not help to distinguish BA analyses from NMDs. BA analyses are more abstract than NMDs with respect to a particular class of detail, but may be less abstract with respect to another class of detail. Second, we argued that the details added into NMDs and missing from in BA explanations need not necessarily concern how the system works. Third, we have proposed reasons for doubting that BA analyses, unlike NMDs, may be considered mechanism sketches. These views are based on a critical examination of claims made by Piccinini and Craver (2011) and Levy and Bechtel (2013). Taken together, they may contribute to further clarifying the relationship between different styles of explanation widely adopted in behavioral sciences, and, therefore, to unifying branches of cognitive science that adopt markedly different theoretical vocabularies.

REFERENCES

Bechtel, W., and Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Scientific Research Strategies.* Princeton: Princeton University Press.

Craver, C. (2007). *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199299317.001.0001

Craver, C. F. (2002). Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philos. Sci.* 69, S83–S97. doi: 10.1086/341836

Craver, C. F. (2006). When mechanistic models explain. *Synthese* 153, 355–376. doi: 10.1007/s11229-006-9097-x

Cummins, R. (1985). *The Nature of Psychological Explanation.* Cambridge, MA: The MIT Press.

Datteri, E., and Laudisa, F. (2012). Model testing, prediction and experimental protocols in neuroscience: a case study. *Stud. Hist. Philos. Biol. Biomed. Sci.* 43, 602–610. doi: 10.1016/j.shpsc.2012.04.001

Kandel, E. R., Schwartz, J. H., Jessell, S., Siegelbaum, A., and Hudspeth, J. (2000). *Principles of Neural Sciences,* 5th Edn. New York: McGraw-Hill.

Levy, A., and Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philos. Sci.* 80, 241–261. doi: 10.1086/670300

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

Rosenblueth, A., and Wiener, N. (1945). The role of models in science. *Philos. Sci.* 12, 316–321. doi: 10.1086/286874

Rosenblueth, A., Wiener, N., and Bigelow, J. (1943). Behavior, purpose and teleology. *Philos. Sci.* 10, 18–24. doi: 10.1086/286788

Wolpert, D. M., Miall, R. C., and Kawato, M. (1998). Internal models in the cerebellum. *Trends Cogn. Sci.* 2, 338–347. doi: 10.1016/S1364-6613(98)01221-2

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biol. Philos.* 25, 287–318. doi: 10.1007/s10539-010-9200-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 March 2014; accepted: 30 April 2014; published online: 22 May 2014.

Citation: Datteri E and Laudisa F (2014) Box-and-arrow explanations need not be more abstract than neuroscientific mechanism descriptions. Front. Psychol. ***5****:464. doi: 10.3389/fpsyg.2014.00464*

This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 03 June 2014
doi: 10.3389/fpsyg.2014.00536

For a science of layered mechanisms: beyond laws, statistics, and correlations

Cristiano Castelfranchi*

Institute of Cognitive Sciences and Technologies, Consiglio Nazionale delle Ricerche, Rome, Italy

Edited by:
Roberto Cordeschi, Sapienza University of Rome, Italy

Reviewed by:
Carlos Eduardo B. De Sousa, Northern Rio de Janeiro State University – Universidade Estadual do Norte Fluminense, Brazil
Russ McBride, University of Utah, USA

****Correspondence:***
Cristiano Castelfranchi, Institute of Cognitive Sciences and Technologies, Consiglio Nazionale delle Ricerche, via San Martino della Battraglia 44, Rome 00185, Italy
e-mail: cristiano.castelfranchi@istc.cnr.it

Two general claims are made in this work. First, we need several different *layers* of "theory," in particular for understanding human behavior. These layers should concern: the cognitive (mental) representations and mechanisms; the neural underlying processes; the evolutionary history and adaptive functions of our cognition and behaviors; the emergent and complex social structures and dynamics, their relation and feedbacks on individual minds and behaviors, and the relationship between internal regulating goals and the external functions/roles of our conduct; the historical and cultural mechanisms shaping our minds and behaviors; the developmental paths. Second, we do not just need "predictions" and "laws" but also "explanations"; that is, we need to identify the mechanisms producing (here-and-now, or diachronically) a given phenomenon. "Laws" are not enough; they are simply descriptive and predictive; we need the "why" and "how." Correlations are not enough (and they are frequently misleading). We need computational models of the processes postulated in our theories[1].

Keywords: reductionism, cognitive architecture, emergence, intentions, functions, computer modeling and simulation, proximate causes

THE NEED FOR EXPLANATION: MAIN ISSUES

We do not just need a "pluralistic" approach (as radically interdisciplinary) but a "layered" theory of our "objects" (Dale, 2008). We need (at least) six layers and axes of theory; not just predictions but explanations, that is, *we need to identify the explicit definition/understanding of the mechanisms producing* (here-and-now or diachronically) *a given phenomenon*. In particular for human behavior we need:

(A) *Modeling the cognitive (mental) mechanisms producing and governing (controlling) our behavior*. That is, we have to explain a given behavior with its "proximate" causes: micro-processes, irreducible to the vocabulary (ontology) of neuro-processes.

(B) *The neural and body implementation of psychological representations and processes*. We should know not only *where* they are located in the brain, but the brain micro-mechanisms and emergent cognitive processes, and *why* they work there.

(C) *The biological evolution of our behavior* (and its "causes": adaptive functions, niche, environmental constraints) and of the mental (cognitive, motivational, affective) mechanisms selected for governing it. Without understanding the "origin," the diachronic causes, we cannot fully explain a phenomenon.

(C1) This requires the understanding of the relation between our genes and our behavior; the dispositions selected by evolution, and how they influence our mental processes and behavior, and how these inherited "programs" interact with experience, learning, and culture.

(C2) This also requires the understanding of the relations between the two kinds of teleology that impinge on us: the internal goals regulating/controlling our action vs. the external functions of our conduct.

(D) *The emergent, collective, self-organizing effects of our behaviors, and their mechanisms and dynamics*; how complexity determines the "social orders." An analytic and dynamic theory of the "invisible hand." Otherwise, we cannot understand societies, etc., as well as the relation between emergent collective phenomena and our intentions and mental representations: *How* is it possible that we "pursue" ends that we do not are aware of and are not among our intentions? We also need to explain how the emergent structure/order feedbacks into, and shapes, our minds and behaviors: not just the "emergence" but also the "immergence" processes.

(E) The historical and cultural evolution – its mechanisms, not just its description and narration – shaping our minds and behaviors, through learning, practices and technologies, norms, and so on. The cultural evolution is not less relevant than the biological one for understanding why we are as we are; and our historical and cultural differences.

(F) We need the modeling of developmental processes, also because some causes of our adult behavior can be found in our personal and relational development (Developmental Cognitive Sciences).

We needs at least all these layers and perspectives (diachronic and evolutional) for *explaining* our behaviors. Laws are not enough. They are simply descriptive and predictive; we need the "why," the "how" (see Cognitive Mechanisms Producing and Controlling Our Behavior & Computational Science for Reconciling "Emergence" with "Cognition"6). Of course, the need for

[1] I would like to thank my friends and collaborators at the ISTC-CNR GoalGroup for discussions and criticisms on these issue, especially Maria Miceli for her help; and Yurij Castelfranchi for useful suggestions and remarks. I also in debt with two anonymous reviewers for their criticisms and suggestions.

explaining (not just predicting and describing) with the underlying devices the observed and observable phenomena, is particularly crucial for the cognitive and behavioral sciences, where the observable phenomena are due to unobservable postulated variables into the minds. But – in my view – is not valid only for human sciences. To be less schematic and more correct, let me say that there are in the natural science "laws" that do really explain in terms of causal underlying "mechanisms" producing the phenomenon and its dynamics. This is the case – in my view – for example of mutation and selection mechanisms explaining Darwinian evolution; although also these mechanisms and laws have to be explained at their micro layer in terms of genes and DNA mechanisms. However, many important "laws" in natural science are not really "explanatory" of the "why" and "how," of the mechanism. For example, the most famous natural law, Newton's gravity, is more descriptive than based on the explanation of the (micro) mechanisms producing/causing attraction. We are still in search of the real causal explanation: the "graviton." At the higher level of course that law predicts but also causally "explain" why something (pears) falls to the ground (from the trees).

No brain map too is enough: it is just cartography, descriptive rather than explicative (see The Neural Implementation of Psychological Representations and Processes).

Correlations are not enough and they are frequently misleading (Concluding Remarks). Theories should be complemented by *models* of the processes that produce and control people's behavior.

(G) This is why a crucial revolution in the behavioral sciences is and will be the computational modeling: the radical "operational" approach. There is no alternative to this especially if one has to model the process at a given micro-layer, and the processes at the macro-layer, and also the emergent (bottom-up) and the immergent (top-down) feedbacks, and how all this works.

I will only focus on issues (A) (C) (D) (G), and in a quite schematic and assertive way also on (B). However, there is a coherence between the central claim of section 5 (point D) on "social" theory (social action and minds are crucial but not enough; we need a theory of the self-organizing macro social order and of its feedback at the micro level) and the seemingly far polemic on (B; brain and mind): layered view is needed because reality is a recursive multi-layered "emergent" complex system: not only we have emergence (and self-organization) from individual to collective, but also from micro neuro-processes and macro functions in brain, and from brain to mental activities, and from cognitive micro-constituents (like beliefs or goals) to complex mental states "gestalts," like an intention (Knowledge–Motivation Commerce), an expectation, or an emotions like hope (Miceli and Castelfranchi, 2014); and so on (A Layered Science for a Layered World).

COGNITIVE MECHANISMS PRODUCING AND CONTROLLING OUR BEHAVIOR

As we said (A), to *explain* a given behavior we need to identify its "proximate" cognitive causes -underlying processes that are irreducible to neuro-processes: representational and functional. Of course, also neural processes are "representational" and based on "functional" notions (like "activation," "inhibition," "connection," etc.), but at a lower micro-level.

In particular, what is needed is a theory of how our behavior is under a "control device"; its mainly goal-governed nature, and how motivations are organized and processed. This is the weakest part of psychology: we know everything about knowledge processing and organization (step by step, all the chapters of a handbook of Cognitive Psychology), but we know very little about motivation and its processing. In particular we should model how our goals (used here as a general term for internal motivational representations during the cybernetic cycle (Miller et al., 1960); including wishes, desires, concerns, intentions, and so on) are processed (activated, chosen, preferred, planned) on the basis of our beliefs, and how we acquire and integrate or revise them. The central mechanism of mind is the "commerce" between goals and beliefs (the two basic families of mental representations). Of course also other "mechanisms" are there; simple reflexes, conditioned reactions, habits, routines, scripts, and so on.

To exemplify the kind of cognitive architecture we should model, let us focus on the last stage of the processing of a goal and on its final package, which regulates *intentional* action (as a specific kind of "behavior").

KNOWLEDGE–MOTIVATION COMMERCE

Intentions are those goals that *actually drive our voluntary actions or are ready/prepared to drive them.* They are not another primitive (like in BDI model inspired by Bratman's theory, e.g., Rao and Georgeff, 1995), a different mental object with respect to goals. They are just a kind of goal: the final stage of a successful goal-processing, which also includes "desires" in the broad sense, with very specific and relevant properties (see also Castelfranchi and Paglieri, 2007). Let's remark that the creation of two distinct "primitives," basic independent notions/objects ("desires" vs. "intentions") is in part due to the wrong choice of adopting (also in accordance with common sense) "desires" as the basic motivational category and source. We criticize this reductive move, and introduce a more general and basic (and not fully common sense) teleonomic notion of "goal." This notion also favors a better unification of goal kinds and a better theory of their structural and dynamic relationships.

In a nutshell (**Figure 1**), in our model an *intention* is a goal that:

(1) Has been activated (by a physiological stimulus, an impulse, or an emotion, or just by a new belief) and processed.
(2) Has been evaluated (beliefs) as not impossible, and not self-realizing or already realized by another agent, and thus *up to us*: we have to act in order to achieve it. An intention is always the intention to "do something" (including inactions). We cannot really have intentions about the actions of other autonomous agents. When we say something like "I have the intention that John goes to Naples" what we actually mean is "I have the intention *to bring it about that* John goes to Naples."
(3) Has been chosen against other possible active and conflicting goals, on the basis of an evaluation (beliefs) of the outcomes and possible costs and we have "decided" to pursue it as *preferable* (greater expected value) to its competitors.
(4) Is consistent with other intentions of ours; a simple goal can be contradictory, inconsistent with other goals, but, once it is chosen, it becomes an intention and has to be coherent

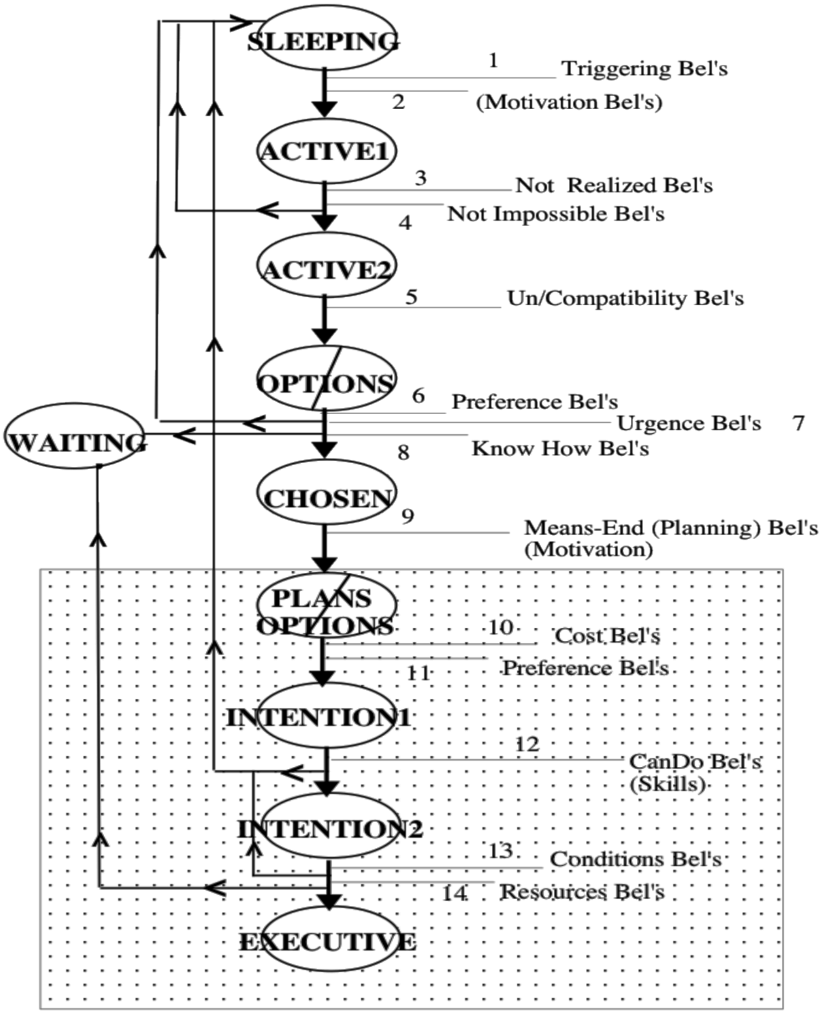


FIGURE 1 | Beliefs in Goal-Processing (Castelfranchi and Paglieri, 2007).

with the other intentions (beliefs about action conditions, resources, and compatibility in the word; Castelfranchi and Paglieri, 2007). Decision-making serves precisely the function of selecting those goals that are feasible and coherent with each other, and allocating resources and planning one's actual behavior.

(5) Implies the agent's beliefs that she knows (or will/can know) how to achieve it, that she is able to perform the needed actions, and that there are or will be the needed conditions for the intention's realization; at least the agent believes that she will be able and in condition to "try."

(6) Being "chosen" implies a *commitment* with ourselves, a mortgage on our future decisions; intentions have priority over new possible competing goals, and are more persistent than the latter (Bratman, 1987).

(7) Is "planned"; we allocate/reserve some resources (means, time, etc.) for it; and we have formulated or decided to formulate a plan consisting of the actions to be performed in order to achieve it. An intention is essentially a two-layer structure: (a) the "intention that," the *aim,* that is, the original goal (for example, to be in Naples tomorrow); (b) the "intention to do," the sub-goals, the planned executive actions (to go to the station, buy the ticket, take the train, etc.). There is no intention without (more or less) specified actions to be performed, and there is no intention without a motivating outcome of such action(s).

(8) Thus an intention is the final product of a successful goal-processing that leads to a goal-driven behavior.

Thus "intention" is not a simple mental object (although outcome of a complex process); it is a *complex configuration* with its *anatomy*: of supporting beliefs and of goals in a means-end relation, and with an impendent commitment.

After a decision to act, an intention is already there even if the concrete actions are not fully specified or are not yet in execution, because some condition for their execution is not currently available. Intentions can be found in two stages:

(a) *Intention "in action,"* that is, guiding the executive intentional action;

(b) *Intention "in agenda"* ("future directed," those more central to the theories of Bratman, Searle, and other), that is, already planned and waiting for some lacking condition for their execution: time, money, skills, etc. For example, I may have the intention to go to Capri next Easter (the implementation of my "desire" of spending Easter in Capri), but now is February 17, and I am not going to Capri or doing anything for that; I have just decided to do so at the right moment; it is already in my "agenda" and binds my resources and future decisions.

I would also say that an "intention" is "conscious," we are aware of our intentions and we "deliberate" about them; however, the problem of unconscious goal-driven behavior is open and quite complex (see Bargh et al., 2001).

During the several steps of its processing, an intention – and its original goal – is supported by those beliefs (on the past, the present, or the future) that are filtering and supporting it. Whenever one of such beliefs changes, there may be a problem for the supported goal, which may be either put in a "waiting room" or abandoned as impossible, already achieved, no longer interesting (because another is deemed to be preferable.), etc. When dealing with *cognitive* agents, in order to change their behavior we have to change their goals (and thus their intentions), but in order to change their goals we have to change their beliefs.

THE NEURAL IMPLEMENTATION OF PSYCHOLOGICAL REPRESENTATIONS AND PROCESSES

A neuroscience of human behavior should *in primis* be the neural modeling of cognitive *mechanisms* and *processes* postulated by the Cognitive Sciences.

Neuroscience shouldn't give us a brain "cartography" of behaviors and feelings: cartography has never been a "science" (just a technique); it explains nothing, it is just description. What we need is the brain/body *implementation* of specific functions and models of elaboration of representations, which determine our conduct.

THE NEUROSIS OF BEHAVIORAL SCIENCES

Neuroscientists shouldn't try to "skip" psychology and its information-processing models of structures and manipulations, for directly connecting brain with behavior (neuro-economics, neuro-aesthetics, neuro-ethics, neuro-politics,. . .). On the contrary they should take the procedural (possibly computational) models of the cognitive sciences and find their neural grounding or – if this proves unfeasible – change them. In fact, a cognitive model that is not grounded in our brain and somatic processes is just wrong, unacceptable. And – on the other side – psychology should provide models of proximate processes; not just correlational "theories," which say nothing on the *mechanisms.*

Actually, there is a minority of approaches that look me rather different and going in a much more promising direction: to analyze the specific "implementation" of *psychological processes and model* in brain functions, processes, and "goals" (Goals vs. Pseudo-goals). Aimed to materialize (they say "embody") cognitive functions in their physical and informational substrate. A very good prototype is for example (Friston et al., 2013) work on the *physical* dynamics in the brain that implement the functions and psychological mechanisms (confidence, expected utility, attainability, inferences, etc.) postulated in decision-making processes.

Nevertheless, in my view, the shortcut temptation I'm pointing on is there, is dominant, and is a misleasing perspective.

The problem is: will neurosciences be able to distinguish, for example, between mere anticipation of benefits or costs, where the expected (and perhaps desirable) result is just predicted, and when this anticipatory representation plays the functional role of (achievement or avoidance) goal? Moreover, expected outcomes that we predict and appreciate/desire are not the same of the expected outcomes that *motivate* our actions: that is, not just additional positive results but those that are *necessary* and *sufficient* for acting.

This is a really crucial distinction (that must be neurologically founded) for a theory of human conduct. Without that it would/will be impossible to distinguish, for example, between:

– Utility-driven vs. value- or norm-driven behavior; or between
– True "altruistic" and non-altruistic pro-social actions.

In fact – in psychological terms – the altruistic nature of an action only depends on the mind-set of the agent. Considering an act as "altruistic" implies a "judgment on mere intent[1]." "Altruistic" is a *subjective* notion, relative to the underlying mental representations (especially the motivational ones); it is not – in human beings – just a behavioral and objective notion. It is not enough that a given conduct is beneficial for Y and costly for X (the agent); even if the benefit is intentional. It is necessary to ascribe to X the motivation to favor Y's wellbeing, rather than some possible expected (external or internal) reward. Thus it would be insufficient to find that these conducts are associated with the activation of a brain area which is related to a "predictive" or anticipatory activity, or to pro-social emotions.

Another example is offered by the neural version of "trust." As Fehr writes: "the rationale for the experiment originates in evidence indicating that oxytocin plays a key role in certain pro-social approach behaviors in non-human mammals. (. . .) Based on the animal literature, Kosfeld et al. (2005), hypothesized that oxytocin might cause humans to exhibit more behavioral trust as measured in the trust game" (Fehr, 2009). In these experiments they also show how oxytocin has *a specific effect on social behavior* because it differently impacts on the trustor and the trustee (only in the first case there is a positive influence). In addition, it is also shown that the trustor's sensitivity to risk is not reduced as a general behavior but it depends on the partner nature (human versus non-human). These are no doubt interesting data. However, the multidimensional and very articulated notion of trust should not be reduced to a generic pro-social attitude and to a particular chemical response or the mere activation of a given brain area. Trust is not a simple, vague, and unitary notion and disposition; it is made of (rather complex) evaluations, expectations, attributions, decisions to rely, sentiments. It should be a componential and analytical psychological model of trust to *drive* the neural

[1] *Beneficium non in eo quod fit aut datur consistit, sed in ipso dantis aut facientis animo*: a benefit consists not in what is done or given, but in the intention of the giver or doer *(Seneca, De Beneficiis Libro I, 6).*

research rather than searching for a simplistic and direct solution, just localistic and correlational (Castelfranchi, 2009).

Analogously, consider norm compliance: will neurosciences be able to distinguish the explicit understanding and processing of a norm and the decision (and reasons) to comply with it, from a merely habitual conforming conduct? And in motivated obedience will neurosciences distinguish between just expected possible sanctions and a decision "motivated" by that avoidance? Will they reduce norms just to the activation of feared punishments or of inhibitory responses? Psychologically speaking, these are very different processes, with quite different socio-political implications.

Finally, we should accept the idea that, in social "games" and scripts, part of the mental attitudes we ascribe to others are not "materially" in their brains. Also mind is an "as if," an "institutional construct." We ascribe certain contents (knowledge, goals,. . .) to others and we act on such as basis, *as if* they were materially there, and this works in our "social pretending": we give them a real, pragmatic, effect, like when we turn pieces of paper into money, by accepting and using them as such (Castelfranchi, 2013). For example, for sure you "know" that $126 + 32 = 158$, or you "know" that Athens is not the capital of Italy, but do you really have this knowledge written in a file of your brain? Not at all! Only after you derived it, not before; however, you implicitly and potentially "know" that and I know that you know and interact with you on such a basis.

Mind is not independent on brain, and in general on a material support of "information processing." Mind is what the brain does but not at its micro level; at the level of macro-functions and complex object (representations). However, "mind" is not only what the brain does. Not only because we might have minds "embodied" in other "machineries" or supports (also at the distributed social interaction level); but because mind is also an "intentional stance" creation, attribution, ascription in order to explain, predict the behavior, and interact with. It is a crucial "instititional" object, even independent of its brain content, like the "value" of money, no longer dependent on gold.

ARBITRARY ASSUMPTIONS IN MIND-BRAIN-BEHAVIOR RELATION

The current views on the relation between psychological processes and their neuro-chemical substratum often betray some questionable assumptions. For example, whereas it is very reasonable to suppose that psychological and support interventions may have an impact on cerebral regulation, at a biological level, this by no means implies that the *origin* of the problem was biological, in terms of a biochemical or neural malfunction.

Mental representations and psychological processes are *per se* IN our brain (if not, where else might they be found?!) and are processes OF our brain. Every construction, acquisition, or elaboration of them just is a neural pattern/process in which our mind materially consists and is implemented[2].

However, to acknowledge this truism does not mean that research at the psychological layer has no longer need to be conducted: psychological notions and models should be neurologically grounded, not "eliminated" (Computational Science for Reconciling "Emergence" with "Cognition"); moreover, one should be aware of the (not just theoretical) risks of biological reductionism and their impact on public opinion; consider for instance the growing tendency of psychiatry to adopt (in theory and in practice) a bio-pharmacological approach, and its problematic consequences at the scientific, social, political, and ethical levels.

Actually, there is a *non sequitur* between the (obvious) idea that dysfunctional/psychopathological (and recovery) processes are *brain processes* and

(i) the assumption that *therefore* their cause *must* be a brain damage, a neural or biochemical dysfunction, a neural *disease*;
(ii) the assumption that *therefore* [even independently of claim (i)] the intervention must necessarily and *directly* be on the brain and its functioning.

To think something is a new state of our brain; to learn something is to modify our brain; to relearn, adjust previous learning, is to modify our brain again. There might have been (for several concurrent factors: internal and external, experiential, relational) a *dysfunctional* learning, dysfunctional thoughts, and the challenge is – through new cognitive and affective experiences and mental elaborations – restructuring the learned representations and processes.

Any change in our conduct or attitudes is/presupposes a change in our minds; any change in our minds is/ presupposes a change in our brains (and bodies). Our brain has been materially "written" by our conduct. In therapeutic, educational or rehabilitation interventions the challenge is to preserve this route, and this view. For changing our brain we do not need to directly act on our brain. Similarly, for producing water we do not need (and it is even worst) to join oxygen and hydrogenous; or for changing genes regulation not necessarily we manipulate genes (epigenetics).

TWO TELEOLOGIES IMPINGING ON HUMAN BEHAVIOR

As for the biological evolution issue (point C), let me just consider a crucial theoretical issue (C2), which is often neglected or mistreated: *the relation between the two kinds of teleology that impinge on us: the internal goals regulating/controlling our action vs. the external functions of our conduct.*

In modern science there are two well-defined teleological frames and notions:

The one provided by *evolutionary approaches*, where it is standard (and correct) to talk in terms of *functions* (adaptive) value, being for something, having a certain finality/end, providing some advantage, etc. In this context "goal" (end, function, finality, etc.) means the "effect" (outcome) that has selected/reproduced and maintained a certain feature or behavior – originally just an accidental effect, an effect among many others, but later, thanks to the loop and positive feedback on its own causes (that is, on the feature or behavior producing it) no longer a mere effect but the *function*, the purpose of that feature, what makes it useful and justifies its reproduction.

- The one provided by *cybernetic control theory* and its postulated cycle, representations, and functions, in which the agent is able

[2] However, also consider other bases of mental entities and process, in term of externalization and distributed cognition, or in radically institutional, conventional terms (as we have just said).

to adjust the world through goal-directed behavior, and to maintain a given desired state of the world (homeostasis).

- Actually, there might be a third teleological/finalistic notion used in several sciences (from medicine to social sciences): the notion of a *function* of X as a *role*, a functional component, an "organ" of a global "system." For example, the function of the heart, or of the kidneys, in our body; or the function of families (or of education or of norms) in a society; or the *function* of a given office in an organization; etc. However, this functionalist and systemic notion has never been well defined and has elicited a lot of problems and criticisms. My view is that this finalistic view is correct, but it is reducible to, and derived from, the previous two kinds of teleology. The organs are either the result of an evolutionary selection – in that they contribute to the fitness and reproduction (maintenance) of that organism – or there is a project, a design, that is, a complex goal in someone else's mind, which imposes particular sub-goals on its parts, components, and tools. Or both.

A serious problem for a (future) science of goals is the fact that these two fundamental teleological notions/mechanisms have never been unified:

(i) Neither conceptually, by looking for a common definition, a conceptual common kernel (for example, in terms of circular causality, feedback, etc.). Do we have and is it possible to have a general, unique notion of "goal" with two sub-kinds (functions vs. psychological goals)?
(ii) Nor by solving the problem of the interaction between the two coexisting forms of finality.

This constitutes a serious obstacle, and reveals a real ignorance gap in contemporary science[3]. For example, as for issue (i), without the aforementioned conceptual unification we cannot have a unitary theory of communication – or a theory of cooperation, of sociality, etc. – in animal and humans. What today are presented as unified theories are just a trick; in fact, those notions – which necessarily require a goal (for example, communication doesn't just require a "reader," it requires a "sender": the information is given on purpose to the receiver/addressee) – are defined in terms of adaptive *functions* when applied to simple animals (like insects), whereas in humans are defined in *intentional* terms. Thus there is no unified notion (and theory) of communication, in that we do not know the common kernel between a *functional* device and an *intentional* device.

Point (ii) is no less problematic. *What is the relationship between the internally represented goals (motivations, and concrete objectives) of an agent regulating its behaviors from the inside, and the adaptive functions that have selected that agent and its behaviors*?

Usually, in purposive, goal-driven agents/systems, the *function* of their conduct, the adaptive result that has to be guaranteed, is *not* internally represented and psychologically pursued; it is not understood and foreseen (**Figure 2**). Of course not all the foreseen outcomes or all the side effects have a *function.*

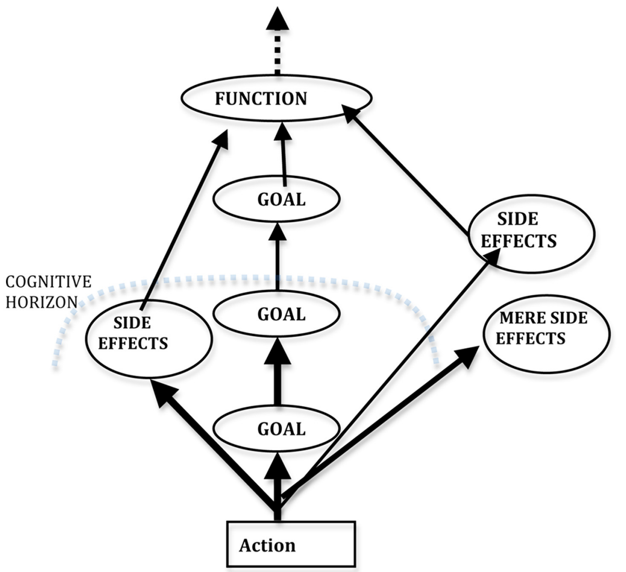


FIGURE 2 | Mental Goals and possible functions.

The internal motivations (and whatever solutions and instrumental goals they generate) may just be sub-goals of the "external" goals of the behavior, of its functions; they are just "cognitive mediators" of the (biological or social) *functions* that would be non-representable and mentally non-computable. For example, only very recently we have discovered why we have to eat, the real functions/effects of our food in our organisms (proteins, carbohydrates, vitamins, etc.); and very few people eat in view of such effects. We eat for hunger or for pleasure or for habit. Analogously, we do not usually make courtship and sex in view of reproduction; we are driven by other internal motives.

Because our behavior may respond to two kinds of teleology – *internal*, driving *goals* (control theory model) vs. *external* selective *functions*, either biological or social (Castelfranchi, 2001) – this is why there might be *conflicts* between one's internal goal and the *function* of one's action/behavior – also considering that we do not necessarily understand and thus intentionally pursue our biological or social functions. We may even act against the functions of our behavior. We may even cut the adaptive connection between our motives and their original functions, for example by deciding to have sex without inseminating or without establishing/maintaining any friendly/affective/supportive relation with our mate. As for social functions, an example of a conflict between our goals and our *role function* could be the goal that B be condemned while I'm his defense attorney.

As for social functions and roles in general, we play them (citizen, consumer, father, pedestrian,. . .) quite blindly; not because they are unconscious, or because just based on reinforcement learning or on mere "habituses" (in Bourdieu's view; see The Emergent, Collective, Self-Organizing Effects of our Behaviors), but because they are external to our minds. In fact, even our intentional and deliberated actions, evaluated on their (visible and

[3] For a deep philosophical and critical discussion of teleologies and the relation with causal explanation in natural sciences, see Larry Wright fundamental work (Wright, 1976). A remarkable attempt to deal with these problems also is Ruth Millikan's work.

conceivable) consequences, may "pursue" collective (good or bad) external "ends" (Castelfranchi, 2001).

For example, if we realize how marketing induces "needs," and deceives and manipulates us, we couldn't play well our most crucial "role" in/for society: the role of "consumers"!

Social functions are *parasitic* to cognition: they establish and maintain themselves *thanks to and through agents' mental representations but not as mental representations: i.e., without being necessarily known or at least intended.* "By pursuing his own interest he [the individual] frequently promotes that of the society more effectually than when he really intends to promote it." (Adam Smith; last sentence of cited paragraph in section "The Emergent, Collective, Self-Organizing Effects of our Behaviors.") However, it is possible, and even frequent, that – following our personal motives – we play our roles in contradiction with the mission and collective utility of our social function.

GOALS VS. PSEUDO-GOALS

It is also very important to disentangle true goals from *pseudo-goals* (Miceli and Castelfranchi, 2012), that is, goals that only seem to be there and to regulate the system and its behavior. However, in fact they are not there as goal mechanisms, they are not represented in the system and "governing" it. They are just functional ways in which the system has been "designed" (by evolution, by learning, by the designer); they are the system's goal-oriented way of working, its operational rules. For example, a real thermostatic system (thermostat, thermometer, room, radiator, boiler, etc.) has been designed in order to reduce naphtha consumption, heat loss, etc. as much as possible. These are (pseudo)goals of the system, which works also in order to guarantee them; but they are not true cybernetic-goals like the set-point of the thermostat. They are not represented, evaluated, and "pursued" by the system action cycle.

Analogously, our minds have been shaped (by natural selection, or culture and learning) in order to have certain working principles and to guarantee certain functions, which are not explicitly represented and intended. It seems (from our behavior) that we have certain goals, but they are not real goals, only pseudo-goals. This is the case, in our view, of some well-known (and badly misunderstood) finalistic notions, like utility maximization, cognitive coherence, and even pleasure. No doubt, we often choose between different possible goals so as to maximize our expected utility, giving precedence/preference to the greater expected value; that is obvious and adaptive. However, this does not mean that we have "the" goal (the unique and monarchic goal) of maximizing our utility, indifferently to the specific contents and goods. On the contrary, we are moved and motivated by specific, qualitative terminal goals of ours (esteem, sex, power, love, etc.), but the *mechanism* that has to manage them has been designed and works so that it maximizes expected utility.

In the same vein, we maintain coherence among our beliefs, and need to avoid and eliminate contradictions. That is why we can reject certain information and do not believe all the data we get (sometimes even what we directly perceive; "we do not believe our eyes," literally); the new data must be plausible, credible, integrable, within the context of our preexisting knowledge; otherwise, we have to revise our previous beliefs on the basis of new (credible) data. This coherence maintenance is frequently completely automatic and routinely. We have mechanisms for coherence check and adjustment. We do not usually have any real intention about the coherence of what we believe. Thus, knowledge coherence is a pseudo-goal of ours, not a real meta-goal guiding meta-actions.

PLEASURE

Similarly, pleasure is not "the" goal of our activity, and the same holds for feeling pleasure (or avoiding feeling pain). "Pleasure" – as a specific and qualitative subjective experience, sensation (not as an empty tautological label for goal satisfaction) – normally is not a goal for us: it is not what we intend to realize/achieve while acting, what move us for performing that behavior. Of course, feeling pleasure or avoiding pain *might* become real goals and intentionally drive our actions: that is basically the mindset of the true hedonist, who acts for pleasure and not for whatever practical consequence his/her action accomplishes. But typically looking for pleasure and avoiding pain are not a unique final goal of ours (another monarchic view of mind and motivation): rather, they act as signals for learning, and they help us learning, among other things, how to generate and evaluate goals.

Those hedonistic philosophies that identify pleasure with motivation, and relate our goal-oriented activity to pleasure motivation, should address the following, evident objections:

i. As a matter of fact, several goals when attained do not give us any pleasure experience; they are just practical results, or consist in the pursuit of (often unpleasant) duties.

ii. If pleasure is so necessary for goal pursuit and motivated activity, why it is not necessary at all in cybernetic models of goal-directed activity and purposive systems? How is it possible to have a clearly finalistic and anticipation-driven mechanism, open to "success" and "failure," without any pleasure? In other terms, *what is the real function and nature of pleasure in a goal-directed system?* Moreover, pleasure seems to be present in nature (both phylogenetically and ontogenetically) well before mentally goal-directed actions. This also suggests that the function of pleasure has to be different; it does not seem to play the role of a goal.

In my view, pleasure is more related to the notion of "reward," of "reinforcement" and learning. Pleasure as an internal reward plays two fundamental roles: it attaches some value to some achieved state, which is important when the system can have more than one of such states, possibly in competition with each other; it signals that a given outcome (perhaps accidental) "deserves" to be pursued, is good, has to become a goal (that state, not the pleasure *per se*). In this view, pleasure is a signal and a learning device for goal creation/discovery and for evaluation. It seems very useful in a system endowed with a "generative" goal mechanism, and which needs different kinds of evaluation, more or less intuitive, fast, based on experience or on biological/inherited "preferences," and not just on reasoning (with its limits, biases, and slowness).

THE EMERGENT, COLLECTIVE, SELF-ORGANIZING EFFECTS OF OUR BEHAVIORS

That is how complexity determines the "social order." Our claim on issue (D) is that we need an analytic and dynamic theory of the "invisible hand," aimed at identifying its underlying mechanisms. Otherwise we cannot understand societies, etc. We also have to understand the relation between the mechanisms regulating the social order and our intentions and mental representations: that is, how – without being understood and explicitly represented – the emergent structure/order feedbacks into and shapes our minds and behaviors; not just the "emergence" but also the "immergence" processes.

The *foundational* issue of the Social Sciences is the micro-macro link, the relation between cognition and individual behavior and social self-organizing phenomena or complex structures and organizations[4]; and institutional actions/phenomena (the two facets of "social order": the "spontaneous" one and the organized or at least institutionalized one; Tummolini and Castelfranchi, 2006). This is the main reason for the existence of the social sciences, what they have to "explain," diachronically and synchronically, in its origin and dynamics.

As remarked by Hayek (1996): "This problem [the problem of the unintentional emergence of order and of spontaneous institutions] is in no way specific to economics. . . *it is without doubt the core problem of the whole of social science.*"

That is also why *Methodological Individualism,* although fundamental or better necessary, is not *sufficient* at all as a framework for explaining social interactions and phenomena (Conte and Castelfranchi, 1995).

Adam Smith's original formulation of "THE problem" is – to me – much deeper and clearer than Hayek's formulation.

The great question is how [the individual] *"which does neither, in general, intend to pursue the public interest, nor is aware of the fact that he is pursuing it, . . . is conducted by an invisible hand to promote an end that is not part of his intention"* (Smith, 1976).

The problem is "*how*" the Invisible Hand does really work; in the end, we should (and could) explain the "mechanism" and its reproductive feedback on the agents' minds and behaviors.

In Smith's view of the "Invisible Hand":

(1) there are intentions and intentional behavior;
(2) some unintended and unaware (long term or complex) effect emerges from this behavior;
(3) but that effect is not just an effect, it is an *end* we pursue, i.e., its orients and controls – in some way – our behavior: we "necessarily operate for" (Smith, ibid.) that result.

Now:

- what does it mean and how is it possible that we promote with our action, we in a sense *pursue* something that is not an intention of ours; that the behavior of an intentional and planning agent be goal-oriented, finalistic, without being intentional?
- in which sense the unintentional effect of our behavior is an "end"?

The real problem is to understand how not only such process coexists with an intentional behavior but also exploits it (Castelfranchi, 2001).

Thus special attention should be devoted not only to the "emergent" bottom-up processes but also to the "immergent" ones: the top-down feedback from emergent phenomena to the agent control-system via learning or through understanding and intending (Conte et al., 2007).

In particular we have to identify which of the macro-level phenomena is or *has to be* mentally represented, understood, and even intended in order to reproduce itself and be effective (as it happens with norms), and to discriminate those that are unintended and blind, and presuppose some form of alienation (like social functions or institutional powers). What we have to explain is also how the Invisible Hand and spontaneous (self-organizing) social order are not so spontaneous and disinterested or optimal for the involved people but do systematically favor powerful agents. What is needed is a criticism to von Hayek's theory (or vulgate) about the spontaneous social order as the best *possible* outcome: the often implicit assumption that *an understanding of the social dynamics, deliberate planning, and intentional pursuuit of non-individual outcomes could never achieve better results.*

How much the epistemic and motivational representations that regulate our intentional conduct are *shaped by* the macro sociological, economic, anthropological, political levels? How the former *are functional to* the latters, not just mere complex effects and consequences?

That is: how could the Spontaneous Order not just *emerge* from our autonomous acts but *maintain and reproduce* itself without actively influencing and reproducing those acts? Which – however – are due to our cognitive representations and processes. Thus it has to shape and reproduce those cognitive mechanisms. *The Invisible Hand works also through and on our minds, by manipulating our mental devices in order to bring out the appropriate (not understood and unintended) outcomes.*

In fact, the problem is not just how a given *equilibrium* (like in simple Games) or *coherence* is achieved and some stable *order* emerges. In order to have a "social order" or an "institution" spontaneous emergence and equilibrium are not enough. They must be "functional," that is self-reproducing by a causal loop.

THE "*COGNITIVE MEDIATORS*" OF SOCIAL PHENOMENA

Social phenomena are due to the agents' behaviors, but. . . the agents' behaviors are due the *mental mechanisms* controlling and (re)producing them.

For example: Our Social Power lies in, consists of, others' *Goals* & *Beliefs*! How do they evaluate us and accept to depend on us. That's why we need Mind-Reading! Not only for adjusting ourselves to the others' interference, but for manipulating and exploiting the others or for helping or punishing them.

Social and cultural phenomena cannot be deeply accounted for without explaining how they work *through the individual agents' minds* (the mental "counterparts" or "mediators" of social phenomena).

Does this mean that social actors fully understand what they do/construct? No, not necessarily.

[4] See for example (with a more traditional approach) Sawyer, 2003 or Prietula et al., 1998; also related with agent-based simulation (Computational Science for Reconciling "Emergence" with "Cognition"). For interdisciplinary and integration based view close to our position, see also (Dale et al., 2013). Although in our perspective goal-directed behavior and intended results and self-organization and spontaneous social order are two complementary and interacting faces of sociality and social theory. We do not think that dynamical system theory is the framework for integrating human interaction "into a broader account." Also because the problem is not the coordination of motor, expressive, and linguistic "inter-action," but of social conventions, scripts, institutions, and collective self-organizing "order." We are more in agreement with the previous Dare's claim about the need for a "pluralistic" approach in Cognitive Science.

That's why we use the term: "mediators": because they are the mental ingredients necessary for producing that social phenomenon or structure without *(necessarily) being the mental representation (understanding or intending) of the social phenomena produced by the behaviors that they determine.*

So, I play and reproduce a "social function" (of father, consumer, the witness of a promise, "public opinion," the follower of a leader, etc.) without necessarily understanding it, but with something specific, corresponding, in my head.

As we said, the problem is social functions impinge not only on our habits and automatic or ritual behaviors, but on our deliberated and intentional actions. Charging only the non-intentional, non- deliberate behaviors with those functional aspects is a simplistic solution: according to such a view, role-playing would just be implemented in "habituses" (Bourdieu and Wacquant, 1992). Thus, when a social actor is consciously deliberating and planning, he would not play a social role, he would be "free." I disagree with such a solution. Social actors play social roles and accomplish their social functions also through their deliberate, intentional actions, however they do so not deliberately. This is precisely the problem to be addressed; and it requires a sophisticated model of intentions. We are back to the issues of (C;Two Teleologies Impinging on Human Behavior).

What is the relationship existing between the social system's goals and the goals internal to its members, which directly regulate their actions?

Are social actors able to understand and represent explicitly in their minds the social system's goals? Or are the goals of the social system simply a projection of the goals of (some of) its members? Or, do the members' goals and plans happen to happily coincide with those of the social system? In other terms: do we intends all the goals we pursue?

Functions establish and maintain themselves thanks to and through agents' mental representations but not as mental representations: i.e., without being known or at least intended.

COMPUTATIONAL SCIENCE FOR RECONCILING "EMERGENCE" WITH "COGNITION"

However, "necessary" doesn't mean "sufficient": Mind is not enough. For "explaining" what is happening at the societal and collective layers we have to model the mind of the actors, but this is insufficient. The "individualistic plus cognitive" approach – even if complemented with "collective intentionality," "joint action," "we intend," etc. – is not sufficient for a social theory and for modeling social processes. Social actors do *not* understand, negotiate, and plan all their collective behavior and cooperative activity. Society is not "team work."

This is the real challenge not only for the behavioral and cognitive sciences but for multi-agent systems and Social AI, and computer-supported societies: *Reconciling Emergence with Cognition.* Emergence and cognition are not incompatible with one another; neither are they two alternative approaches to intelligence and cooperation.

On the one hand, cognition has to be conceived as a level of emergence (from sub-symbolic to symbolic; from objective to subjective; from implicit to explicit).

On the other side, emergent and unaware functional social phenomena (ex. emergent cooperation, and swarm intelligence) should not be modeled only among sub-cognitive agents (Steels, 1990; Mataric, 1992), but also among intelligent agents. In fact, for a theory of cooperation and society among intelligent agents – as we said – *mind is not enough,* and cognition cannot dominate and exhaust social complexity (on that Hayek is right; Hayek, 1967).

This is why a crucial revolution in the behavioral sciences is and will be "computational modeling," with its radical "operational" approach. There is no alternative to this, especially if one has to model at the same time the process at a given micro-layer and the processes at the macro-layer, and also the emergent (bottom-up) and the immergent (top-down) feedbacks, and how all this works.

We need a computational modeling of cognitive representations and manipulation (processing; Cognitive Mechanisms Producing and Controlling Our Behavior) and a computational modeling of their neural implementation and of brain very complex dynamics (The Neural Implementation of Psychological Representations and Processes). The same holds at the social level.

THE *THEORETICAL* MISSION OF SOCIAL SIMULATION

Agent-based computer simulation of social phenomena is the crucial (revolutionary) challenge for the future of behavioral sciences. But why is it so?

As we said, the micro-macro link is the *foundational* issue of the behavioral sciences: they should investigate the relation between cognition and individual behavior, on the one hand, and social self-organizing phenomena or complex structures, organizations, and institutional actions and entities[5] on the other hand. This is the main mission of the social sciences, what they have to "explain," diachronically and synchronically.

No approaches or models for studying this complex phenomenon and eventually understanding its (causal) mechanisms are better than agent-based computer simulation. It is the only approach able to model *at the same time* different layers of processing and their top-down and bottom-up feedbacks and circularity. We can model more or less complex minds (with goals, beliefs, reasoning, decisions, etc., but also emotions, reactions, biases, and perception, learning, etc.) and interaction, dependence networks, group activity, organization, cooperation and competition, norms, roles. And we can observe the internal and external dynamics.

Moreover computer implementation of models provide us a formal validation of the theory predictions, and new experimental data (by simulation).

CONCLUDING REMARKS

What is the correct relation between social and collective human behaviors and the individual mind, and between mind and brain? The answer is: *a well-conceived reductionism,* preserving different (interconnected) ontological layers with their vocabulary (like in chemistry for the notion of "valence" or of "acid").

A LAYERED SCIENCE FOR A LAYERED WORLD

Nature (and, in nature, society) has *different levels of complexity and organization*, with the emergence of macro-level entities,

[5] The two faces of "social order": the "spontaneous" one and the organized or at least institutionalized one.

phenomena and laws, grounded on the entities, properties and mechanisms of the lower layer (micro).

"Reductionism" should not be the "elimination" of the entities, notions, dynamics of a given macro level, considered superfluous once it is explained in terms of their micro-entities. "Reductionism" should be "re-conduction": bringing back and grounding the macro-dynamics on the underlying ones. The theories of the macro layer should be not only *compatible* (non contradictory) with the laws of the micro one; they have to be grounded and derivable. Otherwise they are wrong.

Consider for simplicity the following layers of complexity: let's ground all on physics (particles, atoms, forces, etc.); on top of physics, let's put chemistry, then biology grounded on organic chemistry, then neuroscience, then psychology, then social sciences (economics, sociology, anthropology, politics).

Biology has to be explained in biochemical terms, but we cannot eliminate the notion of "cells" with their new properties and laws, although we have to biochemically know how they "work"[6].

In the same vein, social and collective behavior is due to the conduct of individual actors; but individual action is due to mental representations and processes; therefore the principles of social sciences should be grounded in the underlying mental and behavioral phenomena and laws. However after such a re-conduction is made, we cannot do without such notions as crowd, market, inflation, government, etc.

Science has for example re-conducted chemical "valence" (introduced much before atomic modeling) to atomic properties: particles, their electric charge, etc. To have explained "valence" and how it works doesn't make this notion useless; and we couldn't put aside notions like "acid," "basis," "chemical bonds" although we have a full grounding and understanding of them in atomic terms. The Phlogiston theory has been eliminated, because there was no possible confirmation of the hypothesized processes at the supporting/implementing layer.

Exactly in the same way we have to re-conduct mental representations, functions, and processing to the body and its neural mechanisms and structures; they are just material, informational entities[7]; emergent *functions* of their ground, described in informational/functional terms. If it is not possible to bring them back to their sub-stratum, they are inexistent (like phlogiston); but if they are brought back to their underlying micro-processes, they will not be redundant and eliminable. The psychological notions should be preserved for understanding and explaining "what the brain is *doing*": perceiving, memorizing, retrieving, deciding, pursuing, and so on; at its macro-functional level of activity.

Neural correlates cannot be the right *vocabulary* for explaining human behaviors, just because they are at a micro-level and do not still represent and discriminate the complex "patterns" and their properties and functions (not of their sub-components) at the cognitive and motivational macro-level of working. When we will have the real neural representation of a complex object like a "motivating goal," or an "altruistic intention," or of real "trust attitude" (The Neural Implementation of Psychological Representations and Processes), or a "complex emotion with its appraisal components" like envy, we will have a quasi-complete explanation of it (see previous note), but we will not renounce to that psychological vocabulary; since it holds and works at the functional/informational macro layer. Also because, there are other properties of that entity that are due not to its micro-implementation and mental representation, but to its functions and relations at the macro anthropological, sociological, economic level. A table is a "table," functionally and practically speaking, although it is just a cluster of molecules of a given substance; however, at certain level of use its analysis in physical and material terms is fully irrelevant.

More in general: there are no alternatives to the need for *reading* and *understanding* body in terms of functions, not just in terms of "simple" matter and its physico-chemical processes description.

We look at the kidney as a "filter," at glands in terms of "secretion." Otherwise we do not understand what they do, that is, what they are; which is the sense of the physico-chemical processes that we are describing.

The same obviously holds for our brain (just a body organ). Brain anatomy must be a "political geography," not a "geography" of physical objects/structure: it has to localize the areas of given psychological functions. And brain physiology (activity) – to be understood – requires to be read in terms of active psychic processes. "Mind" is just a functional notion: the high level function of neural activity and patterns; and given its *emergent, functional, informational, semiotic-representational* (and even *institutional*) nature is not "reducible" to brain processes, that just provide its material implementation.

We need a micro-macro theory, a specification of the underlying entities and processes producing given phenomena at the superordinate layer. This is what we call – in strict sense – *Science of "mechanisms."*

THE NEED FOR COMPUTATIONAL MODELING

I claim that there is no alternative to computer modeling. We have to provide not just mathematical or formal models but computational ones, if we want to model the proximate causes of a given phenomenon, and its superficial dynamics; the *underlying "mechanisms" that determine those behaviors.* We also need "synthetic" modeling, that is, the material construction of the modeled entity to show how it actually produces the predicted behaviors/effects in interaction with the environment.

As rightly pointed out by Shieber (2004): "The whole thinking process is still rather mysterious to us, but I believe the attempt to make a thinking machine *will help us greatly in finding out how we think ourselves.*"

Computational/synthetic modeling will be pervasive. It will model any hidden mechanism and "dynamics": from chemical reactions, to DNA, from evolution to psychological mechanisms, to social, economic, historical phenomena. This is the message and the *gift* that ICT and in particular AI has to give to science.

Computational modeling will provide not only "models" and conceptual instruments for the theory, but also *experimental platforms*, new empirical data obtained through simulation, and new hypotheses and predictions. Some experiments will be made possible, which are impossible in "nature," either for practical, social, historical, or moral reasons (demography, urbanistic, etc.)

[6] To say nothing of the "evolutionary theory": a completely new foundation.

[7] However, see note 13.

or for the natural inseparability of some distinguishable mechanisms (for example motivational and emotional mechanisms). This will be crucial both for modeling both proximate causes and diachronic, evolutionary causes.

Computer simulation of neural, cognitive, social mechanisms and dynamics, obviously *is not the only method* for identifying the proximate causes, but the most promising method for (i) their fully *procedural* and *formal* characterization (ii) with the additional advantage of "running" the postulated dynamics and seeing their results (conform or not to predictions), and (iii) to conduct a new precious kind of experiments, in particular useful for complexity and emergent effects. The most promising method/tool, especially for modeling "processes" not just static features (physiology not just anatomy) (iv) at different layers, but interacting; including also the bottom-up and the top-down effects, and the resulting dynamics. It is the only approach able to deal in an integrated way with all these mechanisms.

Moreover, computers are a fundamental device for intelligently *collecting and analyzing relevant data* [from the web, for example "Big Data," and from human behavior in natural conditions (traffic, investments, migration, etc.)]. Also in the sense that major scientific discoveries will be made by computers (able to manage Big Data, to demonstrate theorems, to interpret the laws and mechanisms of that data), but also in the sense of "human" traditional science supported by computational instruments.

In sum, in a few years, science will be "computational"; otherwise it will not be.

"MORE GEOMETRICO DEMOSTRATA"

To be more explicit about psychology status: It is unbelievable that after more than half a century the critical remarks of Wittgenstein on psychology be still valid: "*The confusion and barrenness of psychology* is not to be explained by calling it a "young science"; its state is not comparable with that of physics, for instance, in its beginnings. (. . .) For *in psychology there are experimental methods and conceptual confusion.* (. . .) The existence of the experimental method makes us think we have the means of solving the problems that trouble us; though problem and method pass one another by" (Ludwig Wittgenstein *Investigations* PII p. 232).

In my view, Psychology is one of the few sciences that do not officially have a clearly separated theoretical domain, with its university chairs, conferences, curriculum, . . . like for physics, biology, economics, . . . without any direct experimental activity (in case taking into account and *explaining* the result of their "experimental" discipline; and anticipating of half a century the empirical results, like for Einstein theories).

Psychology – probably because of its "guilty" origin from philosophy, and the consequent inferiority complex to the "hard sciences" – has repressed its theoretical and analytical impulses.

Philosophy is in a sense a party to this somewhat phobic attitude of psychology, because it considers the analytical, formal, theoretical work as a prerogative of its own. However, philosophical contributions, though welcome, cannot replace the theoretical and analytical work that must be internal to psychology.

It is not a matter of "experimental philosophy," it is a matter of "theoretical psychology" (Cognitive Science sometime plays such a role).

We need to have psychological states "more geometrico demostrata."

JUST STATISTICAL LAWS AND CONSTRUCTS, AND PROBABILITY?

However, let us conclude with a query about next future, by following not just optimism of will but pessimism of reason:

Will this analytical and "cognitive mediated" view of social phenomena and dynamics, and of computational Agent-based modeling (we hope for) win?

Not so sure at all: we will attend a short cut of statistics, the impressive power of Big Data, correlations, probability, . . . An already very robust trend. The title of Mayer-Schonberger & Cikier' book is "Big Data: A Revolution That Will Transform How We Live, Work, and Think" and I think that they are absolutely right; but this revolution will be insufficient and even deviating if it will just empower our "predition" capabilities, and will not ground new *theoretical understanding* of the mechanisms and causal processes underlying society and cognition. I worry about Anderson's profecy: "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"; prophecy that of course begin to be based on Big Data! And I care more about scientific aim and frame than about scientific "methods" (See also Anderson, 2008; Harford, 2014).

We are witnessing a growing trend of *predicting without understanding*, without modeling the proximate causes.

Does God play dice?

REFERENCES

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine (Science: Discoveries).* Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Bargh, J., Gollwitzer, P., Lee-Chai, A., Barndollar, K., and Trötschel, R. (2001). The automated will: non conscious activation and pursuit of behavioral goals. *J. Pers. Soc. Psychol.* 81, 1014–1027. doi: 10.1037/0022-3514.81.6.1014

Bourdieu, P., and Wacquant, L. J. (1992). *An Invitation to Reflexive Sociology.* Chicago : University of Chicago Press.

Bratman, M. (1987). *Intention, Plans, and Practical Reason.* Cambridge: Harvard University Press.

Castelfranchi, C. (2001). The theory of social functions. *J. Cogn. Syst. Res.* 2, 5–38. doi: 10.1016/S1389-0417(01)00013-4

Castelfranchi, C. (2009). Review of "neuroeconomics: decision making and the brain" edited by Paul W. Glimcher, Colin Camerer, Russell Poldrack, and Ernst Fehr. *J. Artif. Soc. Soc. Simul.* 12, 6.

Castelfranchi, C. (2013). Minds as social institutions. *Phenomenol. Cogn. Sci.* 12, 335–366.

Castelfranchi, C., and Paglieri, F. (2007). The role of beliefs in goal dynamics: prolegomena to a constructive theory of intentions. *Synthese* 155, 237–263. doi: 10.1007/s11229-006-9156-3

Conte, R., and Castelfranchi, C. (1995). *Cognitive and Social Action.* London: UCL Press.

Conte, R., Andrighetto, G., Campennì, M., and Paolucci, M. (2007), "Emergent and immergent effect in complex social systems," in *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence*, Washington.

Dale, R. (2008). The possibility of a pluralist cognitive science. *J. Exp. Theor. Artif. Intell.* 20, 155–179 doi: 10.1080/09528130802319078

Dale, R., Fusaroli, R., Duran, N. D., and Richardson, D. C. (2013). *The Self-Organization of Human Interaction.* Available at: http://fusaroli.weebly.com/uploads/1/4/7/5/14753784/the_self-organization_of_human_interaction.pdf

Fehr, E. (2009). On the economics and biology of trust. *J. Eur. Econ. Assoc.* 7, 235–266. doi: 10.1162/JEEA.2009.7.2-3.235

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* 7:598. doi: 10.3389/fnhum.2013.00598

Hayek, F. A. (1967). "The result of human action but not of human design," in *Studies in Philosophy, Politics and Economics*, ed. F. A. Hayek (London: Routledge & Kegan), 6–105.

Hayek, F. A. (1996). *Individualism and Economic Order.* Chicago: University of Chicago Press.

Harford, T. (2014). *Big Data: are We Making a Big Mistake?* Available at: http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xfk5fPiX

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. (2005). Oxytocin increases trust in humans. *Nature* 435, 673–676. doi: 10.1038/nature03701

Mataric, M. (1992) "Designing emergent behaviors: from local interactions to collective intelligence," in *Simulation of Adaptive Behavior 2*, eds P. Maes, M. Mataric, J.-A. Meyer, J. Pollack, and S. W. Wilson (Cambridge: MIT Press), 35–44.

Miceli, M., and Castelfranchi, C. (2012) "Coherence of conduct and the self-image," in *Consciousness in Interaction. The Role of the Natural and Social Context in Shaping Consciousness*, eds F. Paglieri and J. Benjamins (Amsterdam: John Benjamins), 151–178.

Miceli, M., and Castelfranchi, C. (2014). *Expectancy, and Emotion.* Oxford University Press.

Miller, G., Galanter, E., and Pribram, K. (1960). *Plans and the Structure of Behavior.* New York: Holt, Rinehart and Winston. doi: 10.1037/10039-000

Prietula, M. J., Carley, K. M., and Gasser, L. (eds). (1998). *Simulating Organizations: Computational Models of Institutions and Groups.* Cambridge, MA: MIT Press.

Rao, A., and Georgeff, M. (1995). "*BDI-agents: from theory to practice*," in *Proceedings of the First International Conference on Multiagent Systems – ICMAS' 95*, ed. V. Lesser (Menlo Park: AAAI Press), 312–319.

Sawyer, R. K. (2003). Artificial societies: multi agent systems and the micro-macro link in sociological theory. *Soc. Methods Res.* 31, 325–363. doi: 10.1177/0049124102239079

Shieber, S. (ed.). (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence.* Cambridge, MA: The MIT Press

Smith, A. (1976). "The wealth of nations (1776)," in *The Glasgow edition of the Works and Correspondence of Adam Smith*, Vol.2b, eds R. H. Campbell and A. S. Skinner (Indianapolis: Liberty Press).

Steels, L. (1990). "Cooperation between distributed agents through self-organization," in *Decentralized AI*, eds Y. Demazeau and J. P. Mueller (Amsterdam: Elsevier).

Tummolini, L., and Castelfranchi, C. (2006). The cognitive and behavioral mediation of institutions: towards an account of institutional actions. *Cogn. Syst. Res.* 7, 307–323. doi: 10.1016/j.cogsys.2005.11.014

Wright, L. (1976). *Teleological Explanations: An Etiological Analysis of Goals and Functions.* Berkeley: University of California Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 January 2014; accepted: 14 May 2014; published online: 03 June 2014.
Citation: Castelfranchi C (2014) For a science of layered mechanisms: beyond laws, statistics, and correlations. Front. Psychol. ***5****:536. doi: 10.3389/fpsyg.2014.00536*
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 02 July 2014
doi: 10.3389/fpsyg.2014.00592

Multi-leveled objects: color as a case study

Liliana Albertazzi[1] and Roberto Poli[2] *

[1] Center for Mind/Brain Sciences and Department of Humanities, University of Trento, Trento, Italy
[2] Department of Sociology and Social Research, University of Trento, Trento, Italy

Edited by:
Roberto Cordeschi, Sapienza University of Rome, Italy

Reviewed by:
Kent Johnson, University of California Irvine, USA
Frederic Fol Leymarie, Goldsmiths, University of London, UK

***Correspondence:**
Roberto Poli, Department of Sociology and Social Research, University of Trento, 26 Verdi street, 38122 Trento, Italy
e-mail: roberto.poli@unitn.it

The paper presents color as a case study for the analysis of phenomena that pertain to several levels of reality and are typically framed by different sciences and disciplines. Color, in fact, is studied by physics, biology, phenomenology, and esthetics, among others. Our thesis is that color is a different entity for each level of reality, and that for this reason color generates different observables in the epistemologies of the different sciences. By analyzing color as a paradigmatic case of an entity naturally spreading over different levels of reality, the paper raises the question as to whether making explicit the usually implicit ontological assumptions embedded within the different observables exploited by the different sciences may eventually clarify some of the difficulties of developing a comprehensive theory of color.

Keywords: color appearances, color models, color systems, levels of reality, phenomenology

INTRODUCTION

What is color? Is it a quality of the phenomenal (subjective) appearance or a property of the physical object? Or both? How are the phenomenal quality and the physical property related to each other? As well known, widely different answers have been provided. Among them are the following: color realism based on the reflectance of light by surfaces, that is, physics; phenomenal objectivism based on sensory-motor contingencies; color subjectivism based on qualia, i.e., atomic color experiences; color as brain product based on the effects of objects within us, from the point of view of the nervous system; and color dispositionalism based on the effects of objects within us, from the point of view of our experiential states (on the different epistemological stances see Byrne and Hilbert, 1997).

Consequently, theories on these various matters have analyzed aspects concerning the nature of stimuli, the workings of the neurons, sensory response, the eminently qualitative nature of phenomenal vision, or aesthetic yield. But on establishing correlations among events at these various levels, one should be careful to avoid epistemologically collapsing their relations of ontological dependence into reductions to either the physical or the neuronal level (Albertazzi, 2006, 2013). In fact, as Köhler wrote

> If someone states that things seen must first be experienced as if they were in the brain, he has not realized that the first part of his statement refers to the visual field as a fact of experience, whilst in the second part, where he uses the expression "the brain," he is speaking of a physical object in physical space. This means that he expects to see parts of visual space localized in relation to parts of physical space, and this notion is entirely impossible.
>
> (Köhler, 1947, p. 213)

The discussion on color continues to suffer from the same shortcomings as denounced by Köhler. It still lacks, for example:

- A categorical classification of the differences among the physical, the neuronal, and the properly psychic (mental) marking the onset of color perceptions.
- A distinction between the color stimuli and subjective color conditions of perceptibility (for example, the assimilative phenomena in color appearances, the role of subjective integrations, the capacity to understand such aspects of colors as the difference between warm and cold or light and heavy colors).
- A precise terminology according to the different levels of analysis, relatively to the different color "observables."
- An explicit correlation between models of color and the specific color observables to which they refer.

The thesis put forward in this study is that *only the framework provided by a properly developed theory of levels of reality can handle the complexity of color perception and color spaces.* The assumption, however, is that the different color observables are not totally independent from one another, in the sense that they are connected by a network of dependencies arising from the different levels of reality.

As a step toward understanding and clarifying the nature of color, this paper suggests verifying whether at least some of the controversial aspects of color understanding depend on different ontological (not epistemological) assumptions. Otherwise stated, we propose to bracket the models' epistemological assumptions as far as is possible in order to better grasp the possible presence of underlying ontological differences.

Color perception is characterized by the presence of different theories based on conflicting primitives (wavelength, neural correlates, color appearances), and parameters (hue, saturation, chroma, brightness, lightness, to mention but a few). Furthermore, a variety of color solids have been proposed as models of the space of colors, including cylindrical, conic, pyramidal, and spherical ones (Billmeyer, 1987). Moreover, even when the different theories adopt the same categories, they define them in different and often conflicting ways.

To make matters worse, even the identification of colors raises major problems: to wit, the color matching procedure, on which most colorimetry is based (Boynton, 1979; Brainard, 1995; Koenderink and van Doorn, 2003; Koenderink, 2010), exploits a severely restricted use of color terms and does not consider what the viewer actually perceives, with the exception of the viewpoint of color differences. The phenomenological aspects of observed colors (Stumpf, 1917; Hering, 1920, 1964; Gelb, 1929; Katz, 1935) remain hidden behind the yes/no responses to just noticeable differences (*jnd* – the units of psychophysical analysis).

The question also arises as how to relate natural language color terms for perceived dimensions of color, i.e., relatively to what kinds of concepts are encoded or not encoded by languages, what are the ontological referents, in what universal and linguistic (or culture-specific) meanings consist, etc. The so-called nature/nurture debate in the field of colors is particularly difficult to address, given the tangled development of the taxonomy of colors over time and in the different languages (Williams, 1976; Dedrick, 1998; Paramei, 2007; Jameson and Komarova, 2009; Rakhilina and Paramei, 2011), the terminology adopted by scientific theories that may define colors according to metrical parameters or differently shaped color spaces, and the question of how the subjective perception of color relates to cross-cultural color naming (Jameson, 2005). Furthermore, many more colors exist perceptively than can be linguistically named (Kuehni, 2007, 2010). The problem is that there are not enough terms to qualify color appearances in simple, precise, and exhaustive terms, if necessary.

Scientific nomenclatures usually adopt severely constrained sets of basic terms and qualifiers. While this may be appropriate for specific uses, such as industrial ones, it is too coarse to capture distinctions that people spontaneously make.

Color nomenclatures usually apply to isolated, uniform patches, or the very simplest configurations of color mondrians. Moreover, special color nomenclatures refer to colors pertaining to particular areas of the entire space of colors. A color nomenclature typically relies on a highly simplified framework based on a small number of qualifiers and their combinations (for instance, "red," "deep red," "dark red," "light red," etc.). These labels are the linguistic translations of numerical expressions. That is, they are operational definitions that do not consider the correlation between perception of color and the linguistic expression that best matches the perception. Perception, depending on different settings, including the physical and the mental, often leads to color terms that do not fit into acknowledged standards.

COLORS AND COLOR TERMS

The CIE (Commission Internationale de l'Eclairage, International Commission of Illumination) definition of color runs as follows: Color (perceived) is the "characteristic of visual perception that can be described by attributes of hue, brightness (or lightness) and colourfulness (or saturation or chroma; see International Lighting Vocabulary [ILV], 2011; Standard CIE S 017/E:2011). A series of notes (http://cie.co.at/index.php?i_ca_id=827) clarify that: "when necessary, to avoid confusion between other meanings of the word, the term "perceived color" may be used (note 1); that "perceived color depends on the spectral distribution of the color stimulus, on the size, shape, structure, and surround of the stimulus area, on the state of adaptation of the observer's visual system, and on the observer's experience of the prevailing and similar situations of observation" (note 2); and that "perceived color may appear in several modes of color appearance. The names for various modes of appearance are intended to distinguish among qualitative and geometric differences of color perceptions. Some of the more important terms of the modes of color appearance are given in 'object color,' 'surface color,' and 'aperture color.' Other modes of color appearance include film color, volume color, illuminant color, body color, and Ganzfeld color. Each of these modes of color appearance may be further qualified by adjectives to describe combinations of color or their spatial and temporal relationships. Other terms that relate to qualitative differences among colors perceived in various modes of color appearance are given in 'luminous color', 'non-luminous color,' 'related colour,' and 'unrelated color"' (note 4).

However, color *terms* can only be linguistic labels of perceived *appearances* of colors, not of physical stimuli because we do not perceive physical stimuli as such. If anything, we perceive colors as a consequence of physical stimulation. Also in this respect, however, the relation between physical stimuli and color appearances is less direct than one might think, or can be taken for granted, given the strong contextual dependence of color appearances (Chevreul, 1839; Albers, 1963). It is our suggestion that grounding color nomenclature on the perceptual experience of subjects provides models more robust than those based on an automatic translation of numerical expressions or geometrical positions in a color space. From this emerges the need to arrive at a robust perceptual definition of color terms.

Natural languages use different types of color terms (Biggam, 2012). Since Berlin and Kay's (1977) seminal book, the literature has drawn on a variety of different methodologies ranging from purely linguistic analyses (Wierzbicka, 2006), to anthropological field researches (MacLaury et al., 2007), mainly with the subministration of Munsell chips[1] (Berlin and Kay, 1977; MacLaury, 1992; Davidoff et al., 1999), and Osgood's semantic differential (Madden et al., 2000). More recently, results from the neurosciences have begun to be used (Kay and McDaniel, 1978; Wuerger et al., 2005). For an extensive review of the different universalist and relativist positions see Da Pos and Albertazzi (2007).

Specifically, as regards basic color terms[2], natural languages segment color appearances according to identifiable patterns. Most languages broadly agree on the prototypicality of linguistic categories for so-called focal colors (Rosch, 1973; Rosch et al., 1976). However, agreement on what aspects are the proper referents of color terms in natural languages is still lacking, because different models refer to different parameters or different aspects of color. Most of the dispute between universalists and relativists on color terms, for example, arises because the exponents of each perspective use concepts of color referring to *different* realities, including stimuli, neural correlates, and color appearances. The usual recourse in these cases to qualifiers such as "'unique," "pure," "primary," "elementary," "basic," "focal," and "prototypical" is widely insufficient, because these qualifiers are themselves far from being univocal. A more systematic framework is needed.

[1] That is, the hues presented in his *Notation* book, see Munsell (1905).

[2] That is, universal color categories assumed to be present in most languages, and in a highly constrained order; (see Berlin and Kay, 1977; Kay and McDaniel, 1978; Kay and Regier, 2006, 2007).

To present one of the customary confusions in addressing colors, it is enlightening to consider the difference between hue and color. Unique (also known as unitary or psychologically primary) colors (Hering, 1920) are colors which do not resemble any other colors, whilst binary, or psychologically mixed colors resemble at least two others. The definition is based on the visual similarity which a color shows, or does not show, with other colors, obtained by pure phenomenological observation. The system of color notation closest to the perception of colors based on their visual similarity is the Natural Color System (NCS, Sivik, 1991). In the NCS, reference to unique hues amounts to reference to yellow, red, blue, and green, while reference to unique colors includes also the achromatic white and black; in fact, from a phenomenological viewpoint, black and white are also perceived as colors. The categories of color and hue are not easily definable, however. Prima facie we might define color as everything that is directly seen, i.e., as the color appearance – defined in CIE as the "aspect of visual perception by which things are recognized by their color" – while hue is the aspect possessed by many colors and which makes them chromatic, distinguishing them from non-chromatic colors. A specific hue is more or less visible in a particular color, in the sense that two colors can be of the same hue: one can see the presence of more red in a highly chromatic color of red hue than in a scantily chromatic color of the same hue (for instance in a whitish pink), although the hue of both is simply red. On the other hand, one can also say that the color most representative of redness is a highly chromatic red. In linguistic terms, talk of a focal color as the most representative color of a category ("the best cues of the category," according to Rosch's prototypical classification; Rosch, 1973; Rosch et al., 1976) makes reference to the color with which the word "red" fits best. In fact, focal color is the color in which one sees what one considers the best red, not a color which belongs to the red hue, which is reddest because it is less blue and less yellow. It is worth noting that the "best" red, differently form "unique" red, can bear cultural connotations as well.

Highly chromatic colors belonging to a bipolar scale between two consecutive hues show different degrees of similarity with the extreme colors of that interval. For instance, the interval defined by the extremes "most chromatic yellow" and "most chromatic red" in which mixed colors appear more or less yellowish or more or less reddish – i.e., are similar to one or the other color in different ways – show different degrees of similarity with the extreme colors of that interval. Linguistically, these intermediate colors can be expressed, for example, in terms of "red and yellow," "saffron," "pumpkin," "orange," "carrot," etc. Not necessarily, however, do these color terms have the same referent, and some may also overlap. For example, a color may appear more or less red either because it is pink or because it is orange: in the former case, the hue is maximally red but little visible (the color is only slightly chromatic); in the latter case, the hue is not very red and the color may be highly chromatic. Consequently, one assesses pink as "very red" because it is only slightly or not at all yellow or blue; and likewise one assesses orange as only slightly red because the "hue" is not very red. However, it seems that one can also make an absolute assessment of how much a color is red, so that orange and pink might be treated equivalently, i.e., the extent to which red (not hue) is visible in them.

The perceptual similarity of the mixed hues to the extremes "red" and "yellow" can be quantified (for instance, halfway in the interval (50–50); or more yellowish than reddish (say, 70–30); and so on. Needless to say, different similarity metrics can be developed.

The problem of the perceptual identification and denomination of colors is particularly complex in the case of mixed colors, such as orange. To be noted is that Berlin and Kay's (1977; see also Kay and Maffi, 1999) eleven basic color terms include both unique colors such as white, black, red, yellow, green, and blue (the first six colors in their list), and mixed colors such as orange. According to Sternheim and Boynton (1966), however, when the orange response category is available in a judgment experiment on the color continuum together with the response categories for red, yellow, and green, orange is used with the lowest reliability, i.e., randomly. When the orange response category is omitted, the hues otherwise associated with orange are completely dispersed into the red and the yellow, though with peaks in either red or yellow. Sternheim and Boynton (1966) therefore conclude that orange is some combination of red and yellow, and that the hues associated with the long wavelength part of the spectrum[3] can be described without the category of orange, and making use of two already known color terms (yellow and red). The superfluous nature of the category "orange" was questioned by Boyton himself in a later study. He interviewed Japanese subjects, who were required to express their degree of agreement on the existence of specific categories related to Berlin and Kay's basic color terms. For 90% of the subjects, the category of orange was well categorized as a salient color, and the category was linguistically expressed by mono-lexemic typical terms different from red and yellow (Uchikawa and Boynton, 1987).

This would imply that, phenomenologically, "orange" lies *exactly midway* between the two pure colors of red and yellow (on the status of "orange" from the point of view of painters, see Garau, 1993). Whenever orange varies from the mid-point between red and yellow, the resulting color is described as yellowish red or reddish yellow, as are the other mixed hues of the same range.

NOMENCLATURES

One of the problems raised by the relationship between color perception and color terms is whether perceptual categorization requires linguistic categories at all. That is: do perceptual categories depend on language, learning and higher cognition, or are they independent from them? Munsell chips are definitely too poor a tool with which to verify this issue experimentally (Lucy and Shweder, 1979; Wierzbicka, 1996, 2006; Lucy, 1997). Testing the possible influence of language on color perception requires a more sophisticated experimental setting, such as having several words available for, say, red, in order to signal different environmental conditions (Green-Armytage, 2006; Winawer et al., 2007). In fact, as we have already noted, there is an indefinite number of color appearances, more than any natural language may encode. Therefore, the question arises as to how to relate natural

[3] The expression in Sternheim and Boynton's paper is unfortunate, because the study refers to "perceived" colors.

language terms for perceived colors and the terminology adopted by scientific theories.

Scientific nomenclatures usually adopt severely constrained sets of basic terms and qualifiers. Four different spaces should be taken into account: (1) The space of colorimetry (to be noted, however, is that there are colorimetric spaces, such as CIELAB and CIECAM (respectively Lab Color Space and Color Appearance Model both published by CIE), that (do not perfectly) represent perceived colors, (2) the physiological space LMS (color space based on human cone cells – LMS stands for L- M- and S-cones) and its derivate DKL (Derrington–Krauskopf–Lennie color space), (3) the space of the linguistic representation of colors, and (4) the space of the subjective perception of colors.

To be noted is that the phenomenological perspective under (4), thus far rarely adopted, is starting to attract attention (Sivik, 1974, 1997; Albertazzi et al., 2012).

For each of these spaces, different theories are customarily developed. Each space requires specific groups of observables. The main issue is that most of the contemporary literature fails to distinguish them as clearly as needed, and therefore has difficulties in addressing the problem of their relations. Since colors, whatever they are, are also, and we would say primarily, a question of perception, one may wonder whether starting from real (i.e., subjective) perceptual experience of color provides information that may escape or remain hidden if one instead starts from other frameworks.

COLOR PRIMITIVES

Color theories use different primitives – and even when they use the same terms, they may define them differently. It is consequently mandatory to be clear about the different terminologies and the ways in which different theories use any given term.

It is generally assumed that color can be described according to the parameters of hue, brightness and saturation (Kuehni, 2003; on measurement see Krantz et al., 1989)[4]. These properties make explicit reference to the relation between a given stimulus (hue correlated with wavelength, brightness correlated with luminance, saturation correlated with purity) and the subsequent subjective experience of a perceiver. On the other hand (see above), it is also often taken for granted that hue, brightness, and saturation are attributes of the color *as perceived*; also taken for granted is what they are correlated with, and what they correspond to; and that they form a 3D space where each of them represents a distinct dimension. These parameters result from innumerable experiments on the physical stimuli, i.e., light spectra, or the power at each wavelength. As it happens, light spectra can be readily measured and characterized by three numbers (the so-called tristimulus values of light). However, the shift is constantly made from properties of light spectra (as measured by the tristimulus values) to properties of the surfaces of seen objects (Wyszecki and Stiles, 1982; Hurlbert, 2013). It is customarily claimed that the tristimulus values specify the response of the standard human eye to the color spectrum. This standard response, however, is far from providing a *general* answer to the ways in which human eyes perceive colors, because the determination of the tristimulus values requires highly specific and severely constrained conditions, i.e., generally isolated colors. To provide an example, visual perception in complex environments where phenomena of contrast and assimilation regularly occur is purposely never taken into consideration: in fact, one of the major self-imposed limits adopted by colorimetric analysis is that it should consider only isolated colors, without taking colors combined with other colors into account (Boynton, 1979).

The problems are compounded because the literature on color defines hue, brightness, and saturation in different, often mutually incompatible, ways. Furthermore, although the distinction among hue, saturation and brightness is correct as far as the properties of light are concerned, it is far from being a "natural" – i.e., "phenomenological" – distinction from the point of view of the perceiver (Stumpf, 1917, p. 8; Katz, 1935). Saturation, for example, is a technical term used to characterize decontextualized light stimuli. According to the CIE definition of saturation, it is "the colourfulness of an area judged in proportion to its brightness" [1136], and in a note it is specified that "For given viewing conditions and at luminance levels within the range of photopic vision, a colour stimulus of a given chromaticity exhibits approximately constant saturation for all luminance levels, except when the brightness is very high." Originally introduced by Helmholtz (1867) – explicitly aware of its arbitrariness from a perceptual point of view – the property of saturation should measure the degree of chromatic content in proportion to the brightness of a color. However, Wyszecki and Stiles (1982) note that the concept of saturation (together with the concept of chroma) is perhaps the most controversial concept in the literature on color appearance. In fact, different systems of color representation differ as to their primitives: for example, one finds chroma in Munsell and Sättigung in Deutsches Institut für Normung (DIN)[5]. Since the definition of saturation changes relatively to the color model adopted, the "'usual" definition of saturation as the "colorfulness" (Hunt, 1986) of a color in relation to its "brightness," or the degree of departure from the gray with the same lightness (all grays having zero saturation), is of little help (Mausfeld, 2003).

Finally, the different meanings of "saturation" or "chroma" are not limited to the different color systems in which they appear. Saturation, in fact, is confused with another phenomenological aspect of color, its insistence or forcefulness, i.e., the fact that a color appears more vivid or brighter in the field (Katz, 1935). These qualities of color carry emotional and affordance-type information like the difference between cold and warm colors (Ou et al., 2004a,b; Xin et al., 2004; Da Pos and Green-Armytage, 2007; Da Pos and Valenti, 2007) and the difference between light and heavy, large and small colors (Arnheim et al., 1954/1974; Itten, 1961), and they concern the theory of the harmonic dimensions of color (Burchett, 2007).

[4] Alternative names for "saturation" are "colorfulness," "intensity," and "purity." Munsell uses instead a different primitive, namely "chroma"; "chromaticness" in NCS. See below for a brief reconstruction of their meaning.

[5] DIN is based on a circle of 24 color-hues, a saturation scale, and a darkness scale as a special parameter for establishing the relative brightness of non-self-illuminating colors (i.e., colors that are illuminated by an external source). See http://www.colorsystem.com/?page_id=948&lang=en.

In past years, "brightness" was sometimes even used as synonymous with "lightness," which fortunately is no longer the case. From a perceptual point of view, "brightness" is an attribute of the light that reaches the eye from a surface, while "lightness" refers to the colors of an object, i.e., it is an attribute of a surface. Lightness is an observable referring to white, understood as the color with the highest lightness (100%). It follows that the lightness of chromatic colors and grays is always less than 100%. Lightness then corresponds to the reflectance of a surface, a property of distal stimulus – that is, a phenomenologically inaccessible property. Using brightness and lightness as synonymous would therefore merge two different observables: an observable of light and an observable of surface. To further compound the confusion, "brightness" may be also used for surfaces, thereby indicating the more or less strong illumination (i.e., *light*) to which they are subject. In this latter case, many technicians prefer to use "luminance" (thereby not referring to the corresponding perception, i.e., brightness). Luminance, in fact, is a psychophysical property pertaining to the stimulus, and not perceivable as such by a perceiver.

Finally, "brightness" is also used for the correlation between the impression of lightness and luminance, where under the same luminance colors of higher saturation appear brighter than colors of low saturation (for example, the Helmholtz, Kohlrausch, and Boswell illusion; see Kaiser, 1985).

FRAMEWORKS OF ANALYSIS

The foregoing discussion has shown how tangled the "scientific" analysis of colors is, and we have provided some evidence about how different some of the presently most widely used theories and approaches are. Some of their differences are due to pragmatic factors such as the needs of the communities using them: for instance, technicians requiring colorimetric data prefer to use either the DIN, the Munsell, the CIELAB or CIECAM02 systems (nowadays with a preference for the last). In one way or another, all the systems need to take account of four different natural systems: physical radiation, physiological elaboration, perception, and language. They differ as to where the focus falls, and therefore in which other system(s) should be kept under control in order to obtain the information they deem relevant. Munsell, NCS, and also OSA-UCS (Optical Society of America, Uniform Color Scale), for example, have a phenomenological base, none of them is primarily focused on physical radiation. Munsell, however, accepting the Fechnerian psychometric law adopts a two-sided understanding of perception, while the NCS adopts and develops a properly phenomenological stance (perception as connected to what appears to awareness), though ruled by psychometric principles.

The Munsell system constrains both psychological and linguistic information: the former by showing individual chips, that is by avoiding contextual influences on color, and the latter by admitting only yes/no answers by the perceiver.

On the other hand, the NCS constrains the neurophysiological base of perception and considers both the source and the neuronal elaboration of the stimuli to be irrelevant. This is not to imply that opponency has no neuronal correlates (Jameson and Hurvich, 1955; MacLeod and van der Twer, 2003; MacLeod, 2010). The problem, however, is that anatomo-physiological substrates cannot explain the phenomenological *qualities* of opponent colors (Valberg, 2001; Kuehni, 2004). As a matter of fact, stimuli for the NCS may arise from any source whatsoever (either "external" or "internal"), and there may be different kinds of them.

By not constraining its phenomenological base, NCS seems to better exploit the richness of both perceptual experience and its linguistic formulation: for example, the relation between warm and cold colors and its linguistic expression (Hård and Sivik, 1981; Da Pos and Valenti, 2007). The very existence of NCS shows that phenomenological observables can produce scientifically exploitable models of color.

The problem remains of making sense of the variety of models. As said, some models are explicitly tailored to the needs of specific communities of users, whilst others are more general in nature. The question however is that all the major models succeed in capturing aspects of the enormously complex problem of color perception. Finding a way to better codify the specific points of view embedded in the various models and systematically coordinate their outcomes may greatly deepen our understanding of colors. Since the discussion has already shown not only that the different models focus on different types of information, but that these types pertain to different sciences (physics, biology, psychology, linguistics – what in a more explicit philosophical parlance becomes the different levels of reality characterizing physical waves, neurophysiological activities, perception, and language – the question arises naturally whether a theory of the levels of reality will indeed be able to clarify and connect, at least to some degree, the different models.

APPROACHING LEVELS OF REALITY

Before presenting some aspects of the theory of levels of reality and their relevance to our topic, some preliminary clarifications on the nature of ontological categories are needed. Needless to say, these clarifications are far from being anything like the presentation of a full-fledged, ontological framework. The economy of the paper forces us to skip issues that a purely philosophical paper would have to address. With these limitations, these clarifications may provide the required anchor points to an ontological framework sufficiently general to clarify scientific models.

We distinguish between categories on the one hand, and individuals on the other, as the entities to which categories refer. Only individuals pertain to the furniture of the world. Categories are not new entities added to the furniture of the world; they are instead principles (or determinations) of the individuals that they categorize. Individuals may be subdivided between *concreta* and *abstracta*, and their categories between real and ideal categories. *Concreta* and real categories pertain to the ontology of real being, *abstracta* and ideal categories to the ontology of ideal (or abstract) being. Universal categories comprise both *concreta* and *abstracta*, real and ideal being. The partial ontology that we are presenting in this paper deals with some aspects of real being only, namely colors.

Moreover, the difference between the nature of categories as principles and the often cumbersome process of their discovery and refinement should never be forgotten. The following quotation aptly summarizes our own understanding of ontological categories:

> "the categories with which ... ontology deals are won neither by a definition of the universal nor through derivation from a formal table of judgment. They are rather gleaned step by step from an observation of existing realities. And since, of course, this method of their discovery does not allow for an absolute criterion of truth, here no more than in any other field of knowledge, it must be added that the procedure of finding and rechecking is a laborious and cumbersome one. Under the limited conditions of human research it requires manifold detours, demands constant corrections, and, like all genuine scholarly work, never comes to an end"
>
> (Hartmann, 1975, p. 13–14).

One of the most difficult problems faced by any ontology is the answer to the following question "What are the individuals to which ontological categories refer?" Two main positions compete; one according to which ontological individuals are only atomic entities, and one which accepts both atomic and molar entities. The former position sees ontological categories as referring to the most elementary components of the universe of discourse (e.g., colors as captured by colorimetry), from which all the other components should derive by composition or other suitable procedures. This is obviously the classic reductionist credo. The alternative vision is more flexible in the sense that it admits a variety of ontological individuals, some of which may work at molar levels of reality (e.g., colors as they appear in the environment, according to phenomena of assimilation and contrast). The main problem facing this alternative vision is that no generally accepted set of intermediate levels arise as the natural candidates from which to start. To compound the difficulty, the various sciences are such that a number of different levels present themselves as "natural" starting points. Selecting any one of them rather than any other is entirely arbitrary. Therefore, there is no saying that the former position is much simpler and (apparently) more effective than the latter. Notwithstanding all the difficulties encountered by the reductionist strategy, many see the reduction to atoms or basic individuals as a perhaps awkward but unavoidable TINA (There Is No Alternative) position. The underlying belief is that the difficulties arising from the reduction to atoms will eventually be solved by more refined strategies, such as new forms of composition. The possibility is usually overcome that even if some individual problem can be reductionistically analyzed, this does not necessarily imply that a generic (that is universal) reductionist strategy is available. Anyway, no patent decision procedure exists to help seriously puzzled scholars to choose between the former and the latter strategy. The unavailability of a proper decision procedure means that in the end the decision depends on a choice that the community of scholars has to take.

Our take on the issue is that the constraint forcing ontological categories to refer to atoms only impoverishes reality in the sense that information is lost and in the end authentic aspects of reality are missed. Instead, an ontological framework acknowledging both atomic and molar categories is both *more general*, in the sense of being able to categorize a wider spectrum of real phenomena, and *more complex*, in the sense of having to address many more problems, such as the ontological nature of the relations between different levels of reality.

This ontological framework systematically distinguishes between "pure" (i.e., "general" or "universal") categories and "domain" (or "level") categories. Keeping in mind this distinction will avert misunderstandings, especially when categories like those of space, time, and causation are introduced.

LEVELS OF REALITY

Today, levels of reality are mostly discussed under the rubrics of "emergence" and "parts and wholes[6]." In fact, the two most obvious strategies with which to approach levels are to divide the world into hierarchies of entities (such as atom–molecule–cell, etc.) or groups of properties (physical, biological, etc.). Not surprisingly, the main distinction among theories of levels of reality closely replicates the divide between entity-based and property-based theories. It is also not surprising that the entity-based theory of levels comes close to part-whole theories, and the property-based theory of levels comes close to type theories. Their merits and demerits notwithstanding, it is worth taking immediate note of an underlying problem: in the above lists of entities/properties, the exact meaning of the concluding "etc." is unclear. Consider the entity-based framework: let us suppose that the series "atom–molecule–cell" will be at some point enlarged by the addition of new entities such as "mind" or "society" (or suitable alternatives). While there are *prima facie* plausible candidates for the relation connecting the items "atom," "molecule," and "cell" (e.g., a part–whole relation), the candidate relations for the new items are remarkably less easy to detect. Similarly, the connections between the properties characterizing "physical" and "biological" types are much simpler (e.g., a subset-set inclusion) than the connections between the properties characterizing the group comprising also "psychological" and "social" types[7].

Of the two main ontological acceptations of entity-based or type-based theories of levels, the former, as said, comes close to the theory of parts and wholes, and the latter to the theory of ontological types. Let us adopt the latter option and understand a level of reality as a group of (ontological) categories (Poli, 2001).

The next step is to distinguish universal categories, those that pertain to the whole of reality, from level categories, those that pertain to one or more levels, but not to all of them. The distinction among physical, biological, psychological, and social types follows naturally. The subsequent step is to specify the relations connecting the levels to each other. Contemporary theories of levels of reality customarily exploit only *one* inter-level relation (e.g., in the form of supervenience). As far as color is concerned, for instance, its phenomenic appearance would be a supervenient product over its physical basis. One of the reasons for rehabilitating Hartmann's theory of levels (see note 6) is that his theory uses *two* different inter-level relations and is therefore able to better distinguish the differences between the physical and the biological levels, on the one hand, and the biological and the psychological levels on the other (Poli, 2006a,b,c, 2007). Provided that the

[6] In the English-speaking world, both strands of analysis have been stimulated by influential papers by Hilary Putnam – notably Oppenheim and Putnam (1958) and Putnam (1961). Since them, an enormous discussion has developed, which cannot be summarized here (for an old but still valuable survey, see Blitz, 1992). However, as important as the discussion in English has been, it is worth noting that some major pre-WWII contributions have never been taken into account, notably those by Nicolai Hartmann. See Hartmann (1940, 1975), Werkmeister (1990), Poli (2012).

[7] Furthermore, beyond or above the distinction between entity-based and type-based theories of levels of reality, other acceptations of levels often intrude, such as notions of levels of organization, complexity or representation.

theory is fully developed and updated to contemporary knowledge, the two relations cover the connections between the physical and the biological levels, on the one hand, and among the biological, psychological, and social (including language and culture) levels on the other (Birren, 1969; Bornstein, 1973). With reference to colors, the two mentioned relations respectively cover stimuli (wavelengths) and their neuro-physiological elaboration (neural correlates), on the one hand, and perceptual modes of appearances of colors (Katz, 1935) and the relations among color terms in natural languages on the other.

As said, the original theory of levels developed by Hartmann is based on *two* different inter-level relations. Leaving universal categories aside, the following two main categorical situations can be distinguished: (a) Beings A and B are categorically different because the categories upon which the former is founded are partially different from the categories upon which the latter is founded, in the sense that the latter is founded on new categories (which implies that the latter includes at least a *novum*, a new category not present in the former); (b) Beings A and B are categorically different because the categories upon which the former is founded and those upon which the latter is founded form two entirely different (disjoint) groups of categories. Following Hartmann, the two relations can be termed respectively relations of super-formation (*Überformung*) and super-position (*Überbauung*; Hartmann, 1940).

Super-formation [the type (a) form of dependence] is weaker than super-position because it includes already actualized categories, those of the level below. Suffice it to consider the super-formation between molecules and cells, i.e., between the physical and the biological levels of reality. In this regard, one can mention that even if organisms are unquestionably more complex than mechanisms, the behavior of organisms complies with the laws of mechanics. On the other hand, the psychological and social levels are different because they are characterized by an interruption in the categorical series and by the onset of new categorical series (relative respectively to the psychological and social levels). The relation between the biological level and the psychological level, on the one hand, and the relation between the psychological level and the social one, on the other, are both relations of super-position. By way of example, the group of categories embedded in psychological entities is different from the group of categories embedded in biological entities. Similarly, the group of categories embedded in social entities is different from the group of categories embedded in biological entities.

When the connecting relation is a relation of super-formation, some categories of the lower level recur in the higher one. Recurring categories interact with the categories of the higher level and are, so to speak, contaminated by them; some of their moments become different. Higher levels are never characterized by recurring categories, however. Each level has its *novum*, the category or group of categories that distinguishes the level from the lower ones. The *novum* does not derive either from the elements of the level or from their synthesis.

Two aspects characterize super-position relations: first, the categories embedded in the entities of the connected levels are different (they are all *nova*); second, a relation of existential dependence links the higher level to the lower one. Most details of the links connecting together the various levels of reality are still unknown, because the various sciences have worked mainly on causal links internal to their regional phenomena[8].

As an observable, color has ramifications into all these different levels of reality and as we have seen the properties of color are different in the different levels.

This is the main reason for at least some of the differences among the different color models. Specifically, the distinction between super-formation and super-position plays a major role. While two different levels related by a relation of super-formation may indeed present the same category, the internal determinants of this category are nevertheless partially different because the category pertains to two different categorical groups: that is to say, it interacts with two different groups of categories. One may say that the category seems to presents an intrinsic ambiguity. We say "seems" because the ambiguity is not embedded in intrinsic features of the category but depends entirely on the observer's shift between different levels of reality (connected by a relation of super-formation). Reading a physical category (the three stimulus codification of a light wave) as a biological category (the three stimulus codification of a neural network) is a case in point.

On the other hand, levels of reality connected by a *super-position* relation present a remarkably different situation. In this latter case – and leaving universal categories aside – the categories defining the two levels are different. In this sense, no ambiguity is likely to arise. Moreover, the two levels are connected by a relation of existential dependence, meaning that the higher level requires the lower one as its existential bearer. Examples from the field of colors are provided by the difference between warm and cold, light and heavy, large and small colors (see Color Primitives above). None of these properties is present in the space of physical radiation. They are authentically phenomenological categories, present only at *that* level of reality. On the other hand, the phenomenological level requires suitable existential bearers – and more than one as a matter of fact: not only the brain as the bearer of the mind, but also the body (because the brain is not an autonomous whole)[9], and the external environment. All of them are required, and all of them are sources of possible perceptual stimulation.

CONCLUSION

As we have seen, color perception is paradigmatic for its complexity, including its ramifications into the physical, the neurophysiological, the linguistic (and cultural) and the phenomenological

[8] The lack of a theory of levels of reality has possibly been the main obstruction against development of the theories needed. Proposals concerning the architecture of levels and their links will improve our understanding of the world and its many dependencies. To mention but one case, the theory of levels paves the way to the claim that there may be different families of times and spaces, each with its own structure. We shall argue that there are numerous types of real times and spaces endowed with structures that may differ greatly from each other. The qualifier *real* is mandatory, since the problem is not the trivial one that different abstract theories of space and time can eventually be, and have been, constructed. We shall treat the general problem of space and time as a problem of chronotopoids (understood jointly, or separated into chronoids and topoids). The guiding intuition is that each stratum of reality comes equipped with its own family of chronotopoids (Poli, 2007; for further details on the theory of levels of reality, see Poli, 1998, 2001, 2006a,b,c, 2009, 2010a,b, 2011a,b, 2012).

[9] Here is where the connection with the theory of levels from the perspective point of the theory of wholes becomes visible.

domains. Some of these ramifications are simpler than others. Not surprisingly, the phenomenological one is the most complex because phenomenic color exists only in the way in which it appears and therefore is a primarily contextual entity deeply influenced by interaction and assimilation (Katz, 1935) and language. The higher complexity of color perception may partly explain the preference shown by many experts for other points of view.

The research hypothesis that we have presented is that the theory of levels may clarify some of these intricacies in the sense of making explicit the ontological references of the various aspects of color, and it may therefore contribute to explaining the concepts of color used in science, phenomenology, and natural language conceptualization.

The analysis has shown that the different models explain color perception by encoding qualities pertaining to different levels of reality, which implies that strictly speaking they model different realities. However, since phenomenic color is essentially a contextual entity, the NCS system seems to be the model closer to color appearances. Further studies may provide additional evidence about whether the explicit connection between a model and the level of reality that it encodes is indeed able to clarify the relations among models themselves (an issue that may be called "ontology as a framework for clarifying science").

ACKNOWLEDGMENT

The authors thank Osvaldo Da Pos for his comments.

REFERENCES

Albers, J. (1963). *Interaction of Color*. New Haven: Yale University Press.

Albertazzi, L. (2006). "Introduction to Visual Spaces," in *Visual Thought. The Depictive Space of Perception*, ed. L. Albertazzi (Amsterdam: Benjamins), 3–33. doi: 10.1075/aicr.67

Albertazzi, L. (2013). "Experimental phenomenology: an introduction," in *Handbook of Experimental Phenomenology: Visual Perception of Shape, Space and Appearance*, ed. L. Albertazzi (Chichester: Wiley-Blackwell), 1–36. doi: 10.1002/9781118329016.ch

Albertazzi, L., Canal, L., Da Pos, O., Micciolo, R., Malfatti, M., and Vescovi, M. (2012). The hue of shapes. *J. Exp. Psychol. Hum. Percept. Perform.* 39, 37–47. doi: 10.1037/a0028816

Arnheim, R. (1954/1974). *Art and Visual Perception: A Psychology of the Creative Eye*. Berkeley: University of California Press.

Berlin, B., and Kay, P. (1977). *Basic Color Terms. Their Universality and Evolution*. Berkeley: University of California Press.

Biggam, C. P. (2012). *The Semantics of Color. A Historical Approach*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139051491

Billmeyer, F. W. (1987). Survey of color order systems. *Color Res. Appl.* 12, 173–186. doi: 10.1002/col.5080120405

Birren, F. (1969). *Light, Colour and Environment: A Thorough Presentation of Facts on the Biological and Psychological Effects of Colour*. New York: Van Nostrand Reinhold.

Blitz, D. (1992). *Emergent Evolution*. Kluwer, Dordrecht. doi: 10.1007/978-94-015-8042-7

Bornstein, M. H. (1973). Color vision and color naming. A psychophysiological hypothesis of cultural difference. *Psychol. Bull.* 80, 257–285. doi: 10.1037/h0034837

Boynton, R. M. (1979). *Human Color Vision*. New York: Holt, Rinehart and Winston.

Brainard, D. H. (1995). "Colorimetry," in *Handbook of Optics:* Vol. 1. *Fundamentals, Techniques, and Design, 26.1-26.54*, 2nd Edn, eds M. Bass, E. Van Stryland, and D. Williams (New York: McGraw-Hill).

Burchett, K. E. (2007). Color harmony attributes. *Color Res. Appl.* 16, 275–278. doi: 10.1002/col.5080160410

Byrne, A., and Hilbert, D. R. (1997). *Readings on Color. The Philosophy of Color*, Vol. 1. Cambridge: MIT Press.

Chevreul, M. E. (1839). *De la loi du Contraste Simultané des Coleurs.* [The Principles of Harmony and Contrast of Colors], ed. F. Birren, English trans. New York: Van Nostrand.

Da Pos, O., and Albertazzi, L. (2007). It is in the nature of color. *Seeing Perceiving* 23, 39–73. doi: 10.1163/187847509X12605137947466

Da Pos, O., and Green-Armytage, P. (2007). Facial expressions, colours and basic emotions. *Colour Des. Creat.* 1, 2, 1–20.

Da Pos, O., and Valenti, V. (2007). "Warm and cold colors," in *AIC Color Science for Industry*, eds Y. Guanrong and X. Haisong (Hangzhou: Color Association of China), 41–44.

Davidoff, J., Davies, I., and Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature* 398, 203–204. doi: 10.1038/18335

Dedrick, D. (1998). *Naming the Rainbow: Colour Language, Colour Science, and Culture*. Dordrecht: Springer. doi: 10.1007/978-94-017-2382-4

Garau, A. (1993). *Le Armonie del Colore* 1984 [Color Harmonies], English trans. Chicago: University of Chicago Press.

Gelb, A. (1929). "Über die 'Farbenkonstanz' der Sehedinge," in *Handbuch der Normalen und Pathologischen Physiologie*, eds A. Bethe, G. von Bergmann, G. Embden, and A. Ellinger. Band 12, *1. Hälfte. Receptionsorgane II* (Berlin: Springer), 594–678.

Green-Armytage, P. (2006). The value of knowledge for color design. *Color Res. Appl.* 31, 253–269. doi: 10.1002/col.20222

Hård, A., and Sivik, L. (1981). NCS-Natural Color System: a Swedish standard for color notation. *Color Res. Appl* 6, 129–138. doi: 10.1002/col.5080060303

Hartmann, N. (1940). *Der Aufbau der Realen Welt: Grundriss der Allgemeinen Kategorienlehre*. Berlin: De Gruyter. doi: 10.1515/9783111442013

Hartmann, N. (1975). *New Ways of Ontology*. Westport: Greenwood Press.

Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*. Hamburg: Voss.

Hering, E. (1964). *Outlines of a Theory of the Light Sense*. Berlin: Springer.

Hering, E. E. (1920). *Zur Lehre vom Lichtsinn* [Outlines of a Theory of the Light Sense], trans. L. M. Hurvich and D. Jameson. Cambridge: Harvard University Press. doi: 10.1007/978-3-662-42443-8

Hunt, R. W. G. (1986). *Measuring Colour*. Chicester: Ellis Horwood.

Hurlbert, A. (2013). "The perceptual quality of color," in *Handbook of Experimental Phenomenology. Visual Perception of Shape, Space and Appearance*, ed. L. Albertazzi (London: Blackwell-Wiley), 369–394. doi: 10.1002/9781118329016.ch15

International Lighting Vocabulary [ILV]. (2011). *Standard CIE S 017/E:2011*. Available at: http://cie.co.at/index.php?i_ca_id=827

Itten, J. (1961). *Kunst der Farbe* [The Art of Color: The Subjective Experience and Objective Rationale of Color]. New York: Van Nostrand Reinhold.

Jameson, D., and Hurvich, L. M. (1955). Some quantitative aspects of an opponent-colors theory. Chromatic responses and spectral saturation. *J. Opt. Soc. Am.* 45, 546–552. doi: 10.1364/JOSA.45.000546

Jameson, K. (2005). Culture and cognition: what is universal about the representation of color experience? *J. Cogn. Cult.* 5, 293–347. doi: 10.1163/156853705774648527

Jameson, K., and Komarova, N. L. (2009). Evolutionary models of color categorization. II. Realistic observer models and population heterogeneity. *J. Opt. Soc. Am.* 26, 1424–1436. doi: 10.1364/JOSAA.26.001424

Kaiser, P. (1985). The Boswell effect (?). *Color Res. Appl.* 10, 186–187. doi: 10.1002/col.5080100308

Katz, D. (1935). *Der Aufbau der Farbwelt* [The Structure of the Color World]. Leipzig: Routledge.

Kay, P., and Maffi, L. (1999). Color appearance and the emergence and evolution of basic color terms. *Am. Anthropol.* 101, 743–760. doi: 10.1525/aa.1999.101.4.743

Kay, P., and McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language* 54, 610–646. doi: 10.1353/lan.1978.0035

Kay, P., and Regier, T. (2006). Resolving the question of color naming universals. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9085–9089. doi: 10.1073/pnas.1532837100

Kay, P., and Regier, T. (2007). Color naming universals: the case of Berinmo. *Cognition* 102, 289–298. doi: 10.1016/j.cognition.2005.12.008

Koenderink, J. J. (2010). *Color for the Sciences*. Cambridge: MIT Press.

Koenderink, J. J., and van Doorn, A. (2003). "Perspectives on color space," in *Color Perception*, eds R. Mausfeld and D. Heyer (Oxford: Oxford University Press), 1–56.

Köhler, W. (1947). *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. New York: Liveright.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1989). *Foundations of Measurement*. Vol. II. *Geometrical, Threshold, and Probabilistic Representations*. San Diego: Academic Press.

Kuehni, R. G. (2003). Theories, technologies, instrumentalities of color: anthropological and historiographic perspectives. *Color Res. Appl.* 28, 231–233. doi: 10.1002/col.10151

Kuehni, R. G. (2004). Variability in unique hue selection: a surprising phenomenon. *Color Res. Appl.* 29, 158–162. doi: 10.1002/col.10237

Kuehni, R. G. (2007). Does the basic color terms discussion suffer from the stimulus error? *J. Cogn. Cult.* 7, 113–117. doi: 10.1163/156853707X171838

Kuehni, R. G. (2010). "Color spaces and color order systems: a primer," in *Color Ontology and Color Science*, eds J. Cohen and M. Matthen (Cambridge: MIT Press), 3–36.

MacLaury, R. E. (1992). From brightness to hue. An explanatory model of color-category evolution. *Curr. Anthropol.* 33, 137–186. doi: 10.1086/204049

MacLaury, R. E., Paramei, G. V., and Dedrick, D. (eds). (2007). *Anthropology of Color*. Amsterdam: John Benjamins Publishing Company.

MacLeod, D. (2010). "Into the neural maze," in *Color Ontology and Color Science*' eds J. Cohen and M. Matthen (Cambridge: MIT Press).

MacLeod, D. I. A., and van der Twer, T. (2003). "The pleistochrome: optimal opponent codes for natural colours," in *Colour Perception: Mind and the Physical World*, eds D. Heyer and R. Mausfeld (New York: Oxford University Press), 155–184.

Madden, T., Hewett, K., and Roth, M. S. (2000). Managing images in different cultures: a cross-national study of color meanings and preferences. *J. Int. Mark.* 2000, 8, 90–107.

Mausfeld, R. (2003). "Competing representations and the mental capacity for conjoint perspectives," in *Inside Pictures: An Interdisciplinary Approach to Picture Perception*, eds H. Hecht, R. Schwartz, and M. Atherton (Cambridge: MIT Press), 17–60.

Munsell, A. H. (1905). *A Color Notation.* Boston: Ellis.

Oppenheim, P., and Putnam, H. (1958). "Unity of science as a working hypothesis," in *Minnesota Studies in Philosophy of Science 2*, eds H. Feigl, M. Scriven, and G. Maxwell (Minneapolis: University of Minnesota Press), 3–37.

Lucy, J. (1997). "The linguistics of colour," in *Color Categories in Thought and Language*, eds C. L. Hardin and L. Maffi (New York: Cambridge University Press), 320–346.

Lucy, J., and Shweder, R. (1979). Whorf and his critics: linguistic and non-linguistic influences on colour naming. *Am. Anthropol.* 81, 581–618. doi: 10.1525/aa.1979.81.3.02a00040

Ou, L. C., Luo, M. R., Woodcock, A., and Wright, A. (2004a). A study of colour emotion and colour preference. Part I: colour emotions for single colours. *Color Res. Appl.* 29, 232–240. doi: 10.1002/col.20010

Ou, L. C., Luo, M. R., Woodcock, A., and Wright, A. (2004b). A study of colour emotion and colour preference. Part II: colour emotions for two-colour combinations. *Color Res. Appl.* 29, 451–457. doi: 10.1002/col.20024

Paramei, G. V. (2007). "Russian 'Blues'. Controversies of basicness," in *The Anthropology of Color*, eds G. MacLaury, D. Dedrick, and G. V. Paramei (Amsterdam: Benjamins Publishing Company), 75–106.

Poli, R. (1998). Levels. *Axiomathes* 9, 197–211. doi: 10.1007/BF02681712

Poli, R. (2001). The basic problem of the theory of levels of reality. *Axiomathes* 12, 261–283. doi: 10.1023/A:1015845217681

Poli, R. (2006a). "First steps in experimental Pphenomenology," in *Artificial Cognition Systems*, eds A. Loula, R. Gudwin, and J. Queiroz (Hershey: Idea Group), 358–386.

Poli, R. (2006b). Levels of reality and the psychological stratum. *Rev. Int. Philos.* 61, 163–180.

Poli, R. (2006c). "The theory of levels of reality and the difference between simple and tangled hierarchies," in *Systemics of Emergence, Research and Development*, eds G. Minari, E. Pessa, and M. Abram (Berlin: Springer), 715–722.

Poli, R. (2007). Three obstructions: forms of causation, chronotopoids, and levels of reality. *Axiomathes* 17, 1–18. doi: 10.1007/s10516-007-9007-y

Poli, R. (2009). Two theories of levels of reality. In Dialogue with Basarab Nicolescu. *Transdisciplin. Sci. Religion* 6, 135–150.

Poli, R. (2010a). "Ontology: the categorial stance," in *Theory and Applications of Ontology.* Vol. 1. *Philosophical Perspectives*, eds R. Poli and J. Seibt (Berlin: Springer), 1–22.

Poli, R. (2010b). Spheres of being and the network of ontological dependences. *Polish J. Philos.* 4, 171–182. doi: 10.5840/pjphil20104223

Poli, R. (2011a). "Hartmann's theory of categories: introductory remarks," in *The Philosophy of Nicolai Hartmann*, eds R. Poli, C. Scognamiglio, and F. Tremblay (Berlin: De Gruyter), 1–32.

Poli, R. (2011b). "Ontology as categorial analysis," in *Classification and Ontology: Formal Approaches and Access to Knowledge*, eds A. Slavic and E. Civallero (Würzburg: Ergon Verlag), 145–157.

Poli, R. (2012). "Nicolai Hartmann," *Stanford Encyclopedia of Philosophy*, Winter 2012 Edn, ed. E. N. Zalta (Stanford, CA: Stanford University). Avalaible at: http://plato.stanford.edu/entries/nicolai-hartmann/

Putnam, H. (1961). "Minds and machines," in *Dimensions of Mind*, ed. S. Hook (New York: Collier), 221–231.

Rakhilina, E. V., and Paramei, G. V. (2011). "Colour terms. Evolution and expansion of taxonomies and constraints," in *New Directions in Colour Studies*, eds C. P. Biggam, C. A. Hough, C. J. Kay, and D. R. Simmons (Amsterdam: Benjamins Publishing Company), 121–132.

Rosch, E. (1973). Natural categories. *Cogn. Psychol.* 4, 328–350. doi: 10.1016/0010-0285(73)90017-0

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cogn. Psychol.* 8, 382–439. doi: 10.1016/0010-0285(76)90013-X

Sivik, L. (1974). *Color Meaning and Perceptual Color Dimensions. A Study of Color Samples.* Göteborg Psychological Reports. 4(1). Göteborg: University of Göteborg.

Sivik, L. (1991). "Cross-cultural studies of color meaning," in *Proceedings of AIC Conference on Color and Light' 91.* Sydney: Colour Society of Australia, 93–96.

Sivik, L. (1997). "Color system for cognitive research," in *Color Categories in Thought and Language*, eds C. L. Hardina and L. Maffi (New York: Cambridge University Press), 163–193. doi: 10.1017/CBO9780511519819.008

Sternheim, C. S., and Boynton, R. M. (1966). Uniqueness of perceived hues investigated with a continuous judgmental technique. *J. Exp. Psychol.* 72, 770–776. doi: 10.1037/h0023739

Stumpf, C. (1917). Zum Gedächnis Lotzes. *Kant Studien* 22, 1–26. doi: 10.1515/kant.1918.22.1-2.1

Uchikawa, K., and Boynton, R. M. (1987). Categorical color perception of Japanese observers: comparison with that of Americans. *Vision Res.* 27, 1825–1833. doi: 10.1016/0042-6989(87)90111-8

Valberg, A. (2001). Unique hues: an old problem for a new generation. *Vision Res.* 41, 1645–1657. doi: 10.1016/S0042-6989(01)00041-4

Werkmeister, W. H. (1990). *Nicolai Hartmann's New Ontology.* Tallahassee: The Florida State University Press.

Wierzbicka, A. (1996). *Semantics. Primes and Universals.* New York: Oxford University Press.

Wierzbicka, A. (2006). "The semantics of colour: a new paradigm," in *Progress in Colour Studies, Volume I. Language and Culture*, eds N. J. Pitchford and C. Biggam (Amsterdam: Benjamins Publishing Company).

Williams, R. (1976). *Keywords: A Vocabulary of Culture and Society.* New York: Oxford University Press.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., and Boroditsky L. (2007). Russian blues reveal effects of language on color discrimination. *Proc. Natl. Acad. Sci. U.S.A.* 104, 7780–7785. doi: 10.1073/pnas.0701644104

Wyszecki, G., and Stiles, W. S. (1982). *Color Science – Concepts and Methods, Quantitative Data and Formulae.* New York: Wiley.

Wuerger, S., Atkinson, P., and Cropper, S. (2005). The cone inputs to the unique-hue mechanism. *Vision Res.* 45, 3210–3223. doi: 10.1016/j.visres.2005.06.016

Xin, J. H., Cheng, K. M., Taylor, G., Sato, T., and Hansuebsai, A. (2004). Cross-regional comparison of colour emotions Part I: Quantitative analysis. *Color Res. Appl.* 29, 451–457. doi: 10.1002/col.20062

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 January 2014; accepted: 27 May 2014; published online: 02 July 2014.
Citation: Albertazzi L and Poli R (2014) Multi-leveled objects: color as a case study. Front. Psychol. ***5****:592. doi: 10.3389/fpsyg.2014.00592*
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 14 July 2014
doi: 10.3389/fpsyg.2014.00668

On agent-based modeling and computational social science

Rosaria Conte and Mario Paolucci*

Laboratory of Agent Based Simulation, Institute of Cognitive Science and Technologies, CNR, Rome, Italy

Edited by:
Roberto Cordeschi, Sapienza University of Rome, Italy

Reviewed by:
Pietro Terna, University of Torino, Italy
Stefano Balietti, ETH Zurich, Switzerland

***Correspondence:**
Mario Paolucci, Laboratory of Agent Based Simulation, Institute of Cognitive Science and Technologies, CNR, via Palestro 32, Rome 00185, Italy
e-mail: mario.paolucci@istc.cnr.it

In the first part of the paper, the field of agent-based modeling (ABM) is discussed focusing on the role of generative theories, aiming at explaining phenomena by growing them. After a brief analysis of the major strengths of the field some crucial weaknesses are analyzed. In particular, the generative power of ABM is found to have been underexploited, as the pressure for simple recipes has prevailed and shadowed the application of rich cognitive models. In the second part of the paper, the renewal of interest for Computational Social Science (CSS) is focused upon, and several of its variants, such as deductive, generative, and complex CSS, are identified and described. In the concluding remarks, an interdisciplinary variant, which takes after ABM, reconciling it with the quantitative one, is proposed as a fundamental requirement for a new program of the CSS.

Keywords: agent-based modeling, computational social science, agent-based simulation, interdisciplinarity, multi-realizability, model building

1. INTRODUCTION

The two decades around the turn of the millennium have seen the rapid advent, and perhaps the premature decline, of a paradigmatic shift in science, represented by agent-based modeling (ABM) and simulation. In this section, after shortly defining what we mean with ABM, we present a short account of its history.

1.1. WHAT AGENT-BASED MODELING IS

What is meant by Agent Based Modeling? Often, this is defined in opposition to Equation-Based Modeling (see for example Dyke Parunak et al., 1998; Cecconi et al., 2010). More specifically, ABM arises at the intersection between agent theory, systems, and architectures, on one hand, and the social sciences, on the other hand. Agents are usually defined (see Conte, 2009) as autonomous systems that operate transitions between states of the world, based on mechanisms and representations somehow incorporated into them.

Under this general definition, the field of agents shows a tremendous variability. Agents vary indeed on several dimensions, which include whether and to what extent they are autonomous, self-interested, sociable, and capable to learn from experience and/or observation. Agents also differ in their level of complexity: according to a classic distinction introduced by Wooldridge and Jennings in their influential work (Wooldridge and Jennings, 1995), agents in a "strong" sense are capable to manipulate and reason upon mental representations; otherwise they are considered agents in a "weak" sense. Another important distinction concerns the way in which mental representations are incorporated: symbolic representations allow an agent to mentally manipulate them in order to reason, plan, take decision, communicate. Sub-symbolic representations are unaware, implicit, based for example on network-like configurations representing the structure of relationships among neurons in cerebral areas, and not liable to purposive manipulation on the side of the agent. Finally, agents vary according to the philosophical or meta-theoretical view their description is based upon. One example is the attempt to model agents on the basis of a personal utility function, on which much work on agents has been done over the past 30–40 years or so, and that has also been criticized as for its micro plausibility (Antunes and Coelho, 2004).

The practice of ABM however did represent a substantial under-exploitation of such wide spectrum of possibilities. De facto, much of the agent models worked out and simulated are totally *ad-hoc*, based on very simple local rules (Epstein, 2006), more or less arbitrarily implemented on a program running on a computer (Gilbert and Troitzsch, 2005). When the program is run, macroscopic effects of the local rules can be observed on the screen, and then be stored, analyzed and possibly visualized in search for emergent phenomena. We will return to the problem of *ad-hoc* rules in section 2.3 below. Such a practice of modeling lends itself well to observe and experiment upon multi-agent worlds or agent societies. These are meant to either reproduce some real-world setting or phenomenon [a typical example is the Anasazi culture simulation (Axtell et al., 2002)], or to build up and observe would-be worlds (Casti, 1997). Such models allow novel theories about abstract social phenomena to be formulated, operationalized, and tested. Examples of this application of ABM abound and are among the best cited works so far worked out in this field.

1.2. AGENT-BASED MODELING IN A HISTORICAL PERSPECTIVE

Conference proceedings, dedicated to the new methodology of ABM and its multiple applications within the social and behavioral areas of science, started to appear in Europe since the early nineties (Gilbert and Doran, 1994; Gilbert and Conte, 1995; Conte et al., 1997). Agent-based models of social phenomena trace back to as early as 1971, when the famous (Schelling, 1971) model of segregation was published in the Journal of Mathematical Sociology. In 2002, the field obtained a major institutional acknowledgment, when the proceedings of a Sackler Colloqium of the National Academy of Sciences, under the title "Adaptive Agents, Intelligence, and Emergent Human Organization: Capturing Complexity through Agent-Based Modeling," held in October 2001, were published on PNAS (Bonabeau, 2002). In that circumstance, ABM was proclaimed as the leading field—we might say the flagship to use a trendy tag - in the renewal of the social, behavioral, and complexity science, which was expected to take place in the years to come. Consecrated by the US scientific institutions, the field was already intensely practiced also, if not primarily, in Europe, where ABM had given rise to a new journal, the Journal of Artificial Societies and Social Simulation (JASSS, founded in 1998), to the first scientific association (ESSA, The European Social Simulation Association, created in 2003), and was at the center of a variety of promotional activities. Soon enough its range of influence extended beyond the two sides of the Atlantic, reaching out to the Pacific area, and giving rise to the PAAA association. At the same time, the NAACSOS association was founded in the US. After some years of fruitful competition, the associations joined in the first World Conference on Social Simulation held in Kyoto in 2006.

At the end of the first decade, however, the ABM leadership seems to be challenged if not decisively weakened by the (re)appearance of a more sober, more encompassing, and less innovative tag, that of Computational Social Science (CSS), of which ABM is a component (see Bankes et al., 2002) for an early insight), and which now candidates itself to replace ABM in the same leading position for the next decade. Evidence of a change of leadership and of a possible coming era for bare CSS, rather than for the more inspiring Generative Social Science proposed by Epstein in 2006, can easily be found in some position papers (e.g., Lazer et al., 2009; Cioffi-Revilla, 2010), books recently appeared (Gilbert, 2010), a new regional association—the CSS Society of the Americas, born on the ashes of the short-lived NAACSOS, and the relative conference held in 2011—and, finally, the objectives of the unsuccessful but groundbreaking EU FET flagship pilot FuturIct (www.futurict.eu).

If history is instructive, the study of signaling is fun. In the era of information overflow, distributed content production, collaborative filtering, crowd sourcing, and so on, emblems are decisive. Tags have a far-reaching but short life. Under the tyranny of PageRank, contents compete in terms of lookups, and these most certainly depend on familiarity, and possibly also on tags appeal. Science makes no difference. It is somewhat surprising when a paradigm shift is signaled by a flat combination of two traditional scientific areas: social sciences and computational science. What is the meaning conveyed by this signal? Does the new label correspond to a new paradigm shift in the social and behavioral sciences, or does it simply meet a kind of marketing need for periodical renewal of names?

This paper presents an attempt to weigh up the impact of ABM and answer the question whether this field is undergoing or not a real decline; whether or not his replacement was timely, necessary, and effective. Next, some current variants of CSS will be compared. Finally, some important requirements for achieving real progresses in the computational study of social phenomena will be identified and discussed.

2. AGENT-BASED MODELING: A BALANCE

Rather than a detailed survey of ABM (for a good example, see Helbing and Balietti, 2011a) this paper presents an attempt to draw a balance of this field, pointing to its main weaknesses and strengths.

2.1. STRENGTH OF AGENT-BASED MODELING

One may wonder what ABM is good for and what are its major strong points. The tricky questions as to when ABM is really needed, whether agent-based models can or cannot be converted into an analytical, equation-based model and to what extent this can be done has been debated at length elsewhere (see for example Epstein, 2006; Cecconi et al., 2010). Nonetheless, ABM remains the only known approach apt to model and reproduce sets of heterogeneous agents interacting and communicating in different ways.

Of course, ABM can only provide a sufficient explanation of the phenomenon of interest, not a necessary one. This feature, which (Epstein, 2006) extensively clarifies and discusses, is also known as *multi-realizability* (Sawyer, 2005), and it is an outstanding property of multilevel systems. A macro-level phenomenon in whatever domain social, natural, mental, etc. of reality, is multi-realizable when it can be implemented in different ways on the lower levels. Inevitably, ABM generates the higher-level effect by following one of the possible generating paths. Even if as many models as generating paths were actually implemented, it would still be difficult, if not impossible, to assess which one among them is effectively implemented in the real world. But, interestingly, this is true also of the target phenomena: an organization can perform its mission independently of the internal structure (consider as an example a project-based structure against a functional one). Social conformity is achieved through a variety of internal mechanisms, e.g., imitation or norm compliance. It is still unclear how disapproval works as a sanction, whether it affects people's decision-making because it activates an expected associated material punishment, or violates the goal of a good (self) esteem. Actually, multi-realizability is a property not only of ABM but also of the real world. In this sense, multi-realizability differs from the more general issue of model underdetermination, as it connects it directly with possible generative paths in reality, an analogy that makes ABM particularly apt to study the equivalence, or possible lack thereof, of structure and mechanisms inside intermediate levels, in the sense of the examples above.

To implement sets of heterogeneous agents in interaction brings about a series of second order advantages: agent societies are (1) operational platforms where theories get converted into falsifiable hypotheses; (2) experimental laboratories where theories get gradually and thoroughly controlled; (3) multilevel worlds where the level of individual units, the agent, is clearly distinct from the macro-level, the system level and unforeseen effects and emergent properties of interaction can be observed.

In short, ABM is an *in nuce* society, which unfolds and actualizes when the model is implemented on a computer program and this runs. In some cases, the effects observed in the computer could not be predicted while modeling and implementing the single units, the individual agents. Hence, the effects of such behavioral units on the whole agent society or parts of it can be observed and investigated. Otherwise stated, ABM allows the interplay between different levels of a social system to be modeled and observed. As shall be seen later, this important property of real-world societies has been insufficiently exploited. The main dynamic investigated by ABM is the way-up of the interaction among the micro and the macro-level. The complementary process, the way-down from the macro-level to the micro-level, has been poorly explored. Closing the loop, however, may require a high level of ABM complexity. Theory-driven, non-*ad-hoc* models of phenomena generated by intelligent behavior may be relatively difficult to calibrate (Heckbert et al., 2010). Difficulties usually increase with the model's level of scale and the number of parameters. One may perceive a trade-off between vertical scaling, i.e., agent complexity, and horizontal scaling, i.e., scenario complexity. Such a trade-off is probably one of the keys for ABM development and leads us straightforward to one of the weak points in the field.

2.2. WEAKNESS OF AGENT-BASED MODELING

Some problems and difficulties in the field of ABM and simulation have been perceived from within the scientific community since long, while others have only recently come to our attention. Since the field's early days, a serious concern of Agent Based modelers and simulators is how to design large-scale agent-based simulations. In its initial applications, agent-based models did not care much about the problem of scale, as they were applied to observe the emergence of patterns from interaction at the microscopic level in artificial scenarios sharing some crucial features of the real-world, but not really aimed to reproduce its details. As soon as the potential of agent-based models became apparent—revealing a great occasion for observing and manipulating *in silico* models of target phenomena in order to acquire a better control, and possibly to optimize intervention—upgrading their level of scale of several orders of magnitude proved necessary. You cannot optimize a system of traffic if you do not manipulate parameters in populations of several millions of agents.

Under the pressure of complex systems science, which is gaining ground in the study of social phenomena (Helbing and Balietti, 2011b), agent-based simulation is increasingly expected to meet a further, and connected, important requirement, i.e., to be fed by massive data in real-time. To answer the problems of scale and real-time simulation, a variety of ICT solutions (parallel and supercomputing infrastructures) are being designed and tested. To deal with this challenge, agent-based simulations were bent to applications needs, such as policy modeling and traffic optimization (Grether et al., 2010), distributed communication over the Internet (Chen, 2009), electricity market (Guerci et al., 2010), financial crisis (Sornette, 2003), epidemics (Pastor-Satorras and Vespignani, 2001). This is not the forum for discussing sophisticated technical solutions (but for a review of techniques to that purpose, the reader might be referred to Paolucci et al., 2013) to the problem of making ABM more apt to the requirements of BigData science. We will instead touch briefly on the question of model equivalence across disciplines and applications.

2.2.1. *Equivalence of models*

Unlike laws of nature, Agent Based models of socio-economic phenomena are countless and not always consistent (see Alfi et al., 2009). Think of the various heuristics and rules of thumb applied in defining microscopic rules for ABM. Most of them generate results at the macroscopic level, which are applied more or less the same narratives or metaphors. Hence, cooperation in Prisoner's Dilemma is found to emerge from a set of heterogeneous strategies, from TIT-FOR-TAT (Axelrod, 1997) to strong reciprocity (Boyd et al., 2003), from image-scoring (Nowak and Sigmund, 1998) to reputation-building (Pinyol et al., 2012), and finally group selection (Di Tosto et al., 2007); social control is found to emerge from ostracism (Xenitidou and Elsenbroich, 2010), but also from partner selection, and finally from gossip (Giardini and Conte, 2012); norms emerge from punishment (Galán and Izquierdo, 2005), which in turn is but a TIT-FOR-TAT strategy, but can also emerge from conditioned preferences (Bicchieri, 2006), and from habituation (Epstein, 2008). Is models' equivalence a major shortcoming of the field, or something social scientists can put up with? What does it depend upon? Is it a necessary or a contingent feature of ABM?

We believe the variety of equivalent agent models in part depends on a property inherent to multi-level systems as complex social systems are. The property in question is the multi-realizability that we have mentioned above. In part, we believe it to be a consequence of the shaky foundations, the poor theoretical justification at the basis of many agent models. This is not equal to finding poorly realistic the model of agent often proposed by current modelers, and asking to improve it toward psychological, cognitive, or sociological plausibility - toward a *seemingly human* agent. What is wrong, in our view, is the procedure for model building and the role of behavioral rules. Let us examine both points with some detail in the next two sections.

2.3. THE ABM RECIPE FOR MODEL BUILDING

A consensus seems to have emerged in ABM on a minimality procedure; that is, models are built by setting up the rules that are minimally required to obtain the macroscopic effect to be described. While minimality might sound obviously inspired by the success of hard sciences, the substantial failure to apply such a minimality procedure to social science is testified by centuries of failed attempts, starting from what had been announced as "Social Physics" in the seventeenth century (for an historical perspective, see Ball, 2002). The reasons for consensus on minimality might be better described with the tools of the sociology of science than rooted in the search of theoretically sound and scientific advances. Indeed, the ABM community, being still relatively small, is subject to issues of disciplinary recognition, with the consequent pressure to publish in a limited number of outlets; and it might still be looking for the right dimension of the contribution - the ideal paper size, as measured in effort invested and soundness of results, could be very different from the "correct" paper size in terms of publication chances. This discrepancy causes a motivational pressure toward minimal (and publishable) models, and hampers research in the much more interesting issue, why minimality seems to fail in the social sciences. We will get back to our intuition on this matter in the conclusions.

Under the rule of minimality, model building is operated (1) a posteriori, based on backward engineering from the effects obtained to the generating rules; (2) *ad-hoc*, so that rules are suggested by the specific results to be obtained; (3) in a rule-oriented rather than agent-oriented approach: what is achieved is a set of rules, rather than an agent view; (4) inspired by the minimal-conditions logic: modeling consists of finding out a set of microscopic rules minimally required to reproduce a given phenomenon of interest. The minimal approach, thus, strongly reduces the validity of ABMs on two separate accounts. On the one hand, theory-based, agent models are implausible caricatures of agent as prescribed by the rationality theory, with a touch of psychological realism in the best possible case. On the other hand, when agent models are not derived from any pre-existing agent theory or vision, whether computational or not, but only by the behavior they are expected to generate (Epstein, 2006), agent models become arbitrary, poorly comparable, competent in highly specific domains of knowledge and disarmingly inapt in any other. It should not come as a surprise if, as a result, a myriad of rather inconsistent agent-based models have been produced over the past 20–30 years or so. Is it possible to find an escape between implausible models and arbitrary ones, or between *ad-hoc* rules and useless ones? Options exist, but are poorly exploited. Why?

2.3.1. From cognitive models. . .

One such option is represented by cognitive agent models, which exist since the late nineties. Their wide range of influence is shown by the popularity of BDI architectures (about 32,700 "BDI agents" cites on Google Scholar retrieved on March, 18th 2013) within and beyond the field of agent systems and theories. Simulation of social phenomena with BDI based models also abound in the literature (about 7060 "BDI social simulation" cites on Google Scholar on March, 18th 2013), and usable platforms to implement them are under consolidation, from Jason (Bordini et al., 2007) to Netlogo extensions (Sakellariou et al., 2008). However, works with this approach receive attention mostly from the computer science community, and are rarely published in main social scientific journals.

Although the rich cognitive models tag appeared since the early nineties (for a recent example see Dignum et al., 2010), the amount of models inspired from it remains negligible. Subsymbolic systems and neural nets did not make much better. Although neural nets and social simulation fare better, relative publications again do not appear in major social scientific journals. Why are cognitive theories on agency, whether symbolic or subsymbolic, so poorly applied in ABM? In part, there are problems of inner validity and calibration. While it is difficult to control the inner validity a complex agent-based model (Cioffi-Revilla, 2002; Windrum et al., 2007), to calibrate it and manipulate parameters values so to reflect a real-world system is hard. To gather data on which the agent model is based upon takes more time and more complex empirical methodology. Therefore, the utility of a complex agent model to simulate the real-world system (i.e., showing that the model's results match the real-world data) is questionable (Crooks et al., 2008). Undoubtedly, these difficulties reduce the interest of cognitively grounded models simulators, although the latter's foundations are much firmer than those of most of the models used. The lesson one might draw is that, like it or not, scientific developments are often due to practical utility more than theoretical soundness. However, the little success of cognitively grounded agent-based models is also due to other factors.

First, unlike other theory-grounded agent models, for example the rational models, cognitive models are not prescriptive. Whereas the theory of rationality is a theory of action, cognitive modeling provides theories of the agent. Hence, the rational agent model fits only apparently better the objectives of ABM and simulation, but it does so only because it allows the modeler to get rid of the tricky part of the modeling, that is, how agents form the goals, the motivations, the preferences, that will be implied in the decisions.

2.3.2. .. to generative models.

Secondly, cognitive modeling is a truly generative theory of behavior, accounting for behavior in terms of the mechanisms that are supposed to operate while producing it. A generative explanation of an observed social phenomenon consists of describing it in terms of the external (environmental and social) and internal (behavioral) mechanisms that generate them, rather than by inferring causes from observed co-variations. This is a vital property of explanation, which cannot easily be realized otherwise. When describing agent behavior by means of other formalisms (logic-based or numeric), we describe behavior from the outside, as perceived by an observer, but do not describe the way it is generated. ABM explains behavior from within, in terms of the mechanisms that are supposed to have generated it, that is, the mechanisms that operate in the agent when s/he behaves one way or another.

Of course, behavior can be explained otherwise. For example, the flight of hawks is wonderfully explained by the mathematical property of logarithmic spiral, such that any tangent from the center of the spiral yields an angle of the same width. Thanks to this property, hawks can keep their preys always in their aim while describing a spiral before pouncing on them. But this explanation is not generative, in the sense that it does not tell us what are the internal mechanisms allowing hawks to fly the way they do. For sure, hawks do not fly based on an understanding of the properties of logarithmic spiral. How can they show the corresponding behavior? The often invoked evolutionary explanation offers poor help: it accounts for behavior in terms of its reproductive advantage. As the spiral-like flight proved advantageous for hawks, those who performed it were able to generate more offspring, while the others extinguished. No generative theory here: it tells us not how hawks produce the behavior in question. We could use the mathematical theory to describe their behavior, and incorporate the mathematical explanation into a set of *ad-hoc* behavioral rules for reproducing it. But neither the mathematical explanation, which describes internal causes, nor the set of *ad-hoc* rules are generative.

Now, a fully generative explanation implies a more general theory of how external causes, including fitness-enhancing effects, get converted into internal reasons (what sometimes are called proximate causes of behavior). Agent-based models are often limited in focus, and not easily compatible with the temporal perspective and the theoretical requirements of a fully generative - in the sense here intended - explanation. Do we always need a generative explanation? Not really, as *ad-hoc* rules sometimes are just all that is needed to explain behavior. This is the case of entirely programmed organisms, and it may even be the case of hawks, as far as we know. Sometimes, instead, you need more. Suppose you want a hawk to learn a new trick with respect to the approach behavior. That is, you, or nature, in the form of new environment - perhaps, but we're letting imagination run wild here, in the form of a prey that develops a counterstrategy to the spiral. Then, immediately how the flight is generated becomes important: how much learned, how much hard wired, and where; in a plastic neuronal connection, or in a fixed relative placement of eye and bone? Suddenly, to reproduce their behavior you would need more than a rigid set of rules; you need to know how it is generated.

Cognitive modeling aims at finding the general mechanisms yielding the wide spectrum of behaviors of relatively autonomous systems. Of course, you don't need such mechanisms to simply reproduce behavior. The more specific the target behavior, the lesser you need a cognitive agent-based model. Since ABM is often used to investigate fairly specific phenomena, either mathematical model or a set of *ad-hoc* rule are preferred over cognitive modeling. But together with cognitive modeling, we also dispense away with truly generative modeling.

2.4. WHY BOTHER WITH GENERATIVE EXPLANATION

One might say, who cares after all? Provided we can reproduce behavior, observe it and make artificial experiments to optimize it, why bother with theory-driven generative modeling? There are several reasons. One is that a truly generative explanation is needed to model complex social dynamics. For universal admission, the dynamics of social entities and phenomena is at least bidirectional if not multidirectional. Entities and properties emerge from the bottom up and retro-act on the systems that have generated them. Current agent-based models instead simulate only emergent properties, i.e., the way up of social dynamics. To mention only a few examples, the ABM literature offers countless models of the emergence of segregation, norms, reciprocity, altruism, cooperation, punishment, conventions, institutions, coalitions, leadership, hierarchies, the modern state. Studies of different types and levels of downward causation are much less frequent (to cite some exceptions, see Gilbert, 2002; Conte et al., 2013). However, how to change self- and other-damaging behaviors (i.e., smoking, over-eating, etc.) was ranked as the fourth most important among the top-ten hard problems the social sciences will have to address in the near future (Giles, 2011).

Agents should not be taken for granted as they change under different types and degrees of social influence. Entities at the macroscopic level affect them and their behavior, and we must understand how this can happen if we want to drive, enforce, or prevent such an influence. This a line of research that presents obvious ethical issues, but at the same time addresses themes so important that social science cannot just leave them alone, or, even worse, desert them to market solutions. For example, at least in some fields, we badly need to know how to reduce or control people's overconfidence, for example in finance, where it so heavily contributed to the last financial crisis (see Akerlof and Shiller, 2010), causing a disruption of global scale; how to change people's bad food habits, which are mainly responsible for highly diffuse diseases as diabetes; how to make low compliant populations to obey the norms, how to increase social trust, reduce hostility toward out-groups, favor communitarian attitudes, and so on. All of these questions might find useful answers based on reality mining. Through Google or Yahoo we may trace people's habits, moods, investment decisions, political views, risk propensity and attitudes toward culture, education and migration. Based on this information, we may drive production, capital movements, business strategies, political decisions, and international cooperation. But we will not be able to suggest effective plans for modifying such behaviors and the underlying mental states, unless we understand the mental dynamics and how this interacts with the social dynamics, and model the cognitive mechanisms that respond to external influence and rule behavioral change. In absence of such theory and model, we will not get to the core of hard problems.

2.5. A MISSED OPPORTUNITY

ABM is a powerful means for investigating the hinge between different domains of reality, including economy, environment, and society: systems' behavior at different levels of scale. It is necessary to explain phenomena pertaining to any domain of reality that is heavily dependent on the behavior of autonomously interactive systems, as was convincingly argued by Epstein. More, ABM is unique for allowing a generative approach to behavioral systems in the sense here defined, and somewhat different from Epstein's, i.e., to describe phenomena in terms of the external and internal mechanisms that produce them.

However, ABM seems to have fulfilled its mission only in part. Its generative capacity has been deployed to a lesser extent than could have been the case. The practice of ABM missed the opportunity it provided: paradoxically, the same principle that led it to a fast popularity, like the KISS principle—i.e., keep it simple, stupid - introduced by Axelrod (1997), and moreover the procedure to find the minimal required conditions to obtain a given phenomenon, do now sentence ABM to a premature end. The KISS principle still drives most of the simulation work: we have performed a check on a whole year of JASSS, a journal that we consider representative of the files. In 2013, JASSS published 49 papers, of which 38 could be classified as simulations (the rest is composed mostly by theoretical papers). Of those, 30 could be considered as following the KISS advice, which makes about the 80% of published papers.

If internal mechanisms are *ad-hoc* and arbitrary, why don't dispense away with them in favor of more powerful quantitative modeling allowing the same phenomena to be accurately predicted? Why bother with agents, if one can apply computational tools to reality mining and platforms to large scale real-world data-driven simulations, and aim at even higher orders of magnitude, enabling us to forecast events at aggregate levels, such as epidemics, climate change, and traffic jams? Couldn't it be the case that a mere quantitative use of computational tools be enough to forecast financial crises, social instability, and even human well-being?

It could be the case, indeed. However, centuries of failed attempts (see the "Social Physics" case mentioned above) make us doubtful. But what is maybe more important, by pursuing this quantitative approach alone, science would have lost a wonderful opportunity: to understand the micro-foundations of phenomena at aggregate levels and how the latter (re)generate them.

3. COMPUTATIONAL SOCIAL SCIENCE

Science, like history, is not a linear process. A decade ago, social, and behavioral science dropped the disciplinary label (Conte, 2002) under the influence of an entirely new field, ABM. In the last couple of years a CSS is being re-proposed. But CSS is being practiced since a couple of decades if not earlier. What is new to the current program?

Computational Social Science (from now on CSS) can be meant in at least three different ways, the deductive, the generative, and the complex one; and it should be made clear which one we are referring to. As these are conceptual, rather than empirical, variants, there is no need to have each of them matching a defined historical example of CSS, since concrete examples are often a mix. Let us characterize variants also with reference to existing programs and try to forecast what their consequences might be.

3.1. THE DEDUCTIVE VARIANT

The second half of the last century is constellated of attempts to apply the theory-building instruments of mathematics and the theory-testing tools of computer science on one side, game theoretic, and logic-based computational models on the other, to describe and explain social phenomena. The latter, in particular, attempted at deducing properties at the macro-level from general assumptions at the micro-level. Expectations á la *homo economicus*, allowed by the theoretical framework, turned out to be wrong, what did not imply that the approach was incorrect, only that it had been based on the wrong assumption, depending on the theoretical and sometimes ideological positions of the authors. What was worse, these position were often left implicit. The deductive variant consists of formulating the mathematical equations that account for the phenomena to be explained. With the support of observation and data gathering, parameters can be assigned their correct values. Although the theoretical framework is often much too simple, the general program scarcely interdisciplinary, and the ambition for social impact mainly based on a rather prescriptive view of micro-level theory, deductive CSS yielded a foundational, general, explanatory theory of social systems. A lesson we should not forget.

3.2. THE GENERATIVE VARIANT

The decline of the rationality paradigm produced several consequences. One of these was a stronger and more interdisciplinary effort to ground computational models on explicit models of the micro-foundations. This led to the advent of the generative variant of ABM, which derives its explanatory vocation and micro-foundational framework from the deductive variant. Unlike it, generative science aims at modeling operational microscopic rules that generate macroscopic phenomena, rather than formulating mathematical equations from which to deduce them. The explanatory vocation is declined in a radically different way: rather than describing a causal process from the outside, the modeler attempts to show the internal rules that initialize it and follow the unfolding of it all the way up to the observed effects.

As argued in the preceding section, however, ABM fulfilled its mission, provide generative theories, to a lesser extent than was expected. If the deductive variant was found to theorize upon fairly abstract phenomena and has often been criticized for its poor predictive capacity, the generative variant did not prove any better at prediction, partly due to problems of validation and calibration.

3.3. THE COMPLEX VARIANT

Inductive computational science is certainly not new (Newell and Simon, 1976). The necessity to combine mathematics and logic with learning, probability, and induction is receiving a growing attention since the early nineties in several computational disciplines such as knowledge representation, reasoning about uncertainty, data mining, and machine learning. Nor is new the use of computational instruments for quantitative social science: it suffices to think of the wide application of statistics package for social scientific research, and by the number of repositories and archives of social scientific data (for example, http://www.data-archive.ac.uk/). However, techniques of data-mining are exercising an even stronger influence on the social sciences. The use of advanced computer technology by social scientists is also shown by sites where freely available web resources are assembled with information on how access social scientific data (see for example, http://guides.lib.wayne.edu/socialsciencesdata), and by funded programs for interfaces between computer and social sciences.

A new impulse to computer-based quantitative social science is coming from the science of complexity, which is now going through a season of deserved popularity. The use of complex systems' methods, models and techniques to economic systems goes back to the nineties (for a rather informative introduction, see Mantegna and Stanley, 2000), and the welcome received by mechanical statistics in the field of economics and finance was such as to encourage its wider application to the rest of the social sciences. The popularity of sociophysics grew even more under the influence of success stories, especially concerning the domain of pedestrians' crowd (Helbing et al., 2000) and that of epidemics (Pastor-Satorras and Vespignani, 2001). In the last few years, a diffuse uncertainty related to globalization, international and cultural conflicts, and the recent financial crisis, led to the necessity to anticipate and manage critical events on the front-stage. Not only stakeholders and policy makers but also, and consequently, research and development funding agencies and evaluators laid emphasis on science as a system of warning, a source of anticipatory information on the performance of aggregate systems, simultaneously triggering and guiding the action of politicians, administrators, and businessmen. But science is more than anticipation. It is first of all explanation. Accurate prediction can do without explaining, especially if it is based on large datasets and sophisticated techniques for extracting knowledge out of them. Science cannot. Of course, explanation may be allowed by statistical analysis. For example, topological properties of complex networks are found among the main factors affecting epidemic dynamics (for a review, see Yang et al., 2007). But this is not always the case. Indeed, this is not the paradigmatic case in those social phenomena in which behavior can be assumed to be irrelevant, or non-influential.

Behavior is irrelevant or non-influential in social dynamics where the implications of the phenomenon in question are social, but its nature is not. To go back to epidemics, the nature of epidemics is biological. The level of reality involved entities belong to does not matter for the observed phenomenon to take place: the nature of entities involved in and target of epidemics matters not. In the spread of epidemics, the difference between human behavior and that of particles in the space does not matter, nor does the difference from carriers and the viruses they carry around. But in other cases, that is, when the nature of behavior matters, accurate statistical analyses of social dynamics can maybe reach predictive power but cannot fully explain what is going on.

As a hypothetical example, suppose we want to know what are the main factors responsible for the dynamics of opinions. Again, current models (Deffuant et al., 2001; Galam, 2002; Hegselmann and Krause, 2002; Castellano et al., 2009), find that the structure of the network of communication affects opinion dynamics. Of course, the source of information also matters; a contrasting source may inhibit the effect of media broadcasting and the process is non-linear: under a given critical level of coverage, the broadcasting message may be inhibited by a "contrarian" opinion spread through word of mouth. Analogously, below a critical level of confidence "contrarian" opinions may reach all agents (Castellano et al., 2009). Social dynamics are often non-linear and typically smolder at some length under the ashes and only subsequently surface in convergent opinions or behaviors. Suppose one predicts the moment(s) at which this will happen in real-world dynamics thanks to statistical analysis and physical models. The question is why it happens. Of course behavior is irrelevant to predict when convergence will occur. But it matters if, for example, we want to affect the process, by shortening or delaying it, or even prevent it; to educate people to a higher autonomy; to favor info-diversity; finally, to convert opinions in something more solid and resistant, like knowledge, and so on.

People withdrawing support from political leaders is a good example of non-linear opinion dynamics. It is unclear when people change their minds and turn down their leaders. The destiny of a popular (and often populist) figure is often decided upon in a very short time. Today, those who enjoyed the favor of their followers until yesterday, may suddenly lose popularity and fall in disgrace, what is again a matter of threshold: after a certain level of spreading, and perceived spreading, agents are led to modify their opinions, what probably reveals an interesting effect of shared representations about shared opinions on one's confidence level. Possibly, such a lowering confidence leads agents to be more eager to change opinions.

However, the circuit may be completely different: agents may resist pressure to change opinions despite contrasting evidence for reasons of cognitive dissonance. The more the contrasting evidence they gather, the higher the dissonance. To reduce it, they try to ignore evidence that is less costly than change opinion, which imply dropping the previous commitment and making a new one. As the perceived distance from others' opinions increases, however, agents either hide their opinions or must defend them openly. If they choose the latter strategy, they may even end up by accepting to form part of a minority. If they take the former option they cannot get along with deception too long as cognitive dissonance increases. Consequently, they accept others' opinions as own, and are likelier to convert them in open behaviors to convince others and themselves about the solidity of their new opinions. Both routes imply critical thresholds for totally different reasons. To act efficaciously on this process, we must be clear what is actually going on. Confidence has different implications from cognitive dissonance and self-deception. To increase confidence may lead to higher stability in the former case, but not in the latter.

To sum up, to model social dynamics without taking into account the internal (cognitive) dynamics of the entities involved in a social phenomenon does not prevent accurate predictions of critical events and changes. It may even allow to find out factors responsible for such events and changes, and this is the case with dynamics for which the social nature of behavior is irrelevant. To understand internal dynamics is crucial instead whenever we need not only to anticipate but also to understand events for which behavior is relevant. Model the internal dynamics of events is necessary not only for scientific reasons but also for guiding intervention.

4. TOWARD A NEW INTERDISCIPLINARY FOUNDATION FOR CSS

The program for CSS needs clarification. Why would such a program be necessary, if we practice CSS since at least a couple of decades? Of course one might say that we need to introduce a new Curriculum at the academic level, and that to do this implies to form a new, cohesive, scientific community, form associations, give visibility to this new Curriculum, strengthen the academic, editorial, and political power of the underlying community etc. However, the reason for a program on CSS is not only political but also scientific. As seen so far, there are different variants of CSS and to take a pluralistic approach to it may be considered wise. CSS could be seen today as a larger umbrella under which different approaches might coexist and somehow feel legitimate. Hence, generative ABM might be practiced by a subset of social scientists, while others might prefer a purely quantitative approach, based on data-mining and numerical simulation, and still others might continue to formulate abstract theories of social action in elegant equations and deduce their macro-level consequences.

The main thesis of this paper is that such a *multidisciplinary* program for CSS would be another lost occasion for science. It would but result either in tight but essentially useless theories, or in accurate predictions of poorly understood social phenomena. In the best possible case, mathematicians will go on citing one another in fairly close circuits of beautiful minds, physicists will find new phenomena affected by the properties of scale-free networks, and social scientists will give up generative ABM in the desperate attempt to produce competitive quantitative social science and get reasonably high scientific scores.

An interesting, innovative program in CSS can only be *interdisciplinary*. Why and where is the difference? The reason lies in the necessity to take advantage from the different modeling methods and techniques to both understand and predict the same phenomena! The difference of interdisciplinary from multidisciplinary CSS consists not only of a convergent investigation of the same phenomena from different perspectives and involving different competencies, what would already be a step ahead with regard to current practice, but in a more radical process aimed at multilevel and modular modeling. Such a type of modeling would allow to describe the dynamics of given phenomena at aggregate levels based on large datasets, find out the criticalities thanks to complex dynamic systems models, make hypotheses about the behavioral dynamics when this is relevant, use ABM to check internal consistency and observe the resulting states at the aggregate level, apply cross-methodological experimental methods to validate the hypotheses against real-world data, update data-mining methods, models, and probability distribution models to newly acquired knowledge, and use mathematical equations when possible to close the number of states resulting at the aggregate level.

An interdisciplinary endeavor like this certainly points out some new challenges: not only to extract knowledge from larger and larger datasets, not only develop simulators that scale up of several orders of magnitude, or feed simulation and data-mining with online real-data, not only to develop supercomputing infrastructures and systems to transfer data to supercomputing platforms, but also develop simulation platforms that scale up both in terms of systems' dimensions and in terms of levels of complexity. We need to account for large-scale systems as well as more complex entities. We need to apply simulation methods to understand the social and the mental dynamics and to describe their interrelationships. Last, but not least, we need incentives that are compatible with such an endeavor—publication-wise and career-wise. This is a challenge for a program on CSS that deserves attention and investment. CSS ought to accept it, or another occasion will be lost for founding a novel, integrated, interdisciplinary, falsifiable science of society helping us to solve transformative and foundational problems.

AUTHOR CONTRIBUTIONS

Rosaria Conte and Mario Paolucci elaborated together the ideas presented in the paper. Rosaria Conte drafted the first version. Mario Paolucci substantially revised it for important intellectual content.

FUNDING

This paper benefits from the stimulating discussions in the *FuturICT* FET Flagship Pilot Project (EU FP7/2007-2013, g.a. no. 284709). The publication of this work is funded by the PRISMA project (PON04a2 A), within the Italian National Program for Research and Innovation (Programma Operativo Nazionale Ricerca e Competitività 2007–2013).

ACKNOWLEDGMENTS

We thank Neil Yorke-Smith, Gustavo Nardin, and Federico Cecconi for encouragement and suggestions. We are also thankful to anonymous reviewers who helped us to tune the paper and helped to clarify some obscure passages.

REFERENCES

Akerlof, G. A., and Shiller, R. J. (2010). *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*. Princeton, NJ: Princeton University Press.

Alfi, V., Cristelli, M., Pietronero, L., and Zaccaria, A. (2009). Minimal agent based model for financial markets I. *Eur. Phys. J. B Condens. Matter Complex Syst.* 67, 385–397. doi: 10.1140/epjb/e2009-00028-4

Antunes, L., and Coelho, H. (2004). "On how to conduct experimental research with self-motivated agents," in *Regulated Agent-Based Social Systems*, Volume 2934 of *Lecture Notes in Computer Science*, eds G. Lindemann, D. Moldt, and M. Paolucci (Berlin; Heidelberg: Springer), 31–47.

Axelrod, R. (1997). *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration, 1st Edn*. Princeton, NJ: Princeton University Press.

Axtell, R. L., Epstein, J. M., Dean, J. S., Gumerman, G. J., Swedlund, A. C., Harburger, J., et al. (2002). Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7275–7279. doi: 10.1073/pnas.092080799

Ball, P. (2002). The physical modelling of society: a historical perspective. *Phys. A Stat. Mech. Appl.* 314, 1–14. doi: 10.1016/S0378-4371(02)01042-7

Bankes, S., Lempert, R., and Popper, S. (2002). Making computational social science effective: epistemology, methodology, and technology. *Soc. Sci. Comput. Rev.* 20, 377–388. doi: 10.1177/089443902237317

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge, MA: Cambridge University Press.

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci. U.S.A.* 99(Suppl. 3), 7280–7287. doi: 10.1073/pnas.082080899

Bordini, R. H., Hübner, J. F., and Wooldridge, M. (2007). *Programming Multi-Agent Systems in AgentSpeak Using Jason*. Chichester: John Wiley & Sons.

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proc. Natl. Acad. Sci. U.S.A.* 100, 3531–3535. doi: 10.1073/pnas.0630443100

Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Rev. Modern Phys.* 81, 591. doi: 10.1103/RevModPhys.81.591

Casti, J. L. (1997). *Would-Be Worlds: How Simulation is Changing the Frontiers of Science*. New York, NY: John Wiley and Sons.

Cecconi, F., Campenni, M., Andrighetto, G., and Conte, R. (2010). What do agent-based and equation-based modelling tell us about social conventions: the clash between ABM and EBM in a congestion game framework. *JASSS* 13.

Chen, D. (2009). "A grid aware large scale agent-based simulation system," in *Quantitative Quality of Service for Grid Computing*, eds L. Wang, J. Chen, and W. Jie (Hershey, PA: IGI Global), 299–319.

Cioffi-Revilla, C. (2002). Invariance and universality in social agent-based simulations. *Proc. Natl. Acad. Sci. U.S.A.* 99(Suppl. 3), 7314–7316. doi: 10.1073/pnas.082081499

Cioffi-Revilla, C. (2010). Computational social science. *Wires Comp. Stat.* 2, 259–271. doi: 10.1002/wics.95

Conte, R. (2002). Agent-based modeling for understanding social intelligence. *Proc. Natl. Acad. Sci. U.S.A.* 99(Suppl. 3), 7189–7190. doi: 10.1073/pnas.072078999

Conte, R. (2009). "From simulation to theory (and backward)," in *Epistemological Aspects of Computer Simulation in the Social Sciences, Chapter From Simulation to Theory (and Backward)*, ed F. Squazzoni (Berlin; Heidelberg: Springer-Verlag), 29–47.

Conte, R., Andrighetto, G., and Campennl, M. (2013). *Minding Norms: Mechanisms and Dynamics of Social Order in Agent Societies*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199812677.001.0001

Conte, R., Hegselmann, R., and Terna, P. (eds.). (1997). *Simulating Social Phenomena*. Vol. 456. Berlin: Springer-Verlag. doi: 10.1007/978-3-662-03366-1

Crooks, A., Castle, C., and Batty, M. (2008). Key challenges in agent-based modelling for geo-spatial simulation. *Comput. Environ. Urban Syst.* 32, 417–430. doi: 10.1016/j.compenvurbsys.2008.09.004

Deffuant, G., Neau, D., Amblard, F., and Weisbuch, G. (2001). Mixing beliefs among interacting agents. *Adv. Complex Syst.* 3, 87–98. doi: 10.1142/S0219525900000078

Dignum, V., Tranier, J., and Dignum, F. (2010). Simulation of intermediation using rich cognitive agents. *Simul. Model. Pract. Theor.* 18, 1526–1536. doi: 10.1016/j.simpat.2010.05.011

Di Tosto, G., Paolucci, M., and Conte, R. (2007). Altruism among simple and smart vampires. *Int. J. Coop. Inf. Syst.* 16, 51–66. doi: 10.1142/S0218843007001561

Dyke Parunak, H., Savit, R., and Riolo, R. L. (1998). "Agent-based modeling vs. equation-based modeling: a case study and users' guide," in *Multi-Agent Systems and Agent-Based Simulation*, Volume 1534 of *Lecture Notes in Computer Science, Chapter 2*, eds J. S. Sichman, R. Conte, and N. Gilbert (Berlin; Heidelberg: Springer Berlin Heidelberg), 10–25.

Epstein, J. M. (2006). *Generative Social Science: Studies in Agent-Based Computational Modeling (Princeton Studies in Complexity)*. Princeton, NJ: Princeton University Press.

Epstein, J. M. (2008). Why model? *JASSS* 11.

Galam, S. (2002). Minority opinion spreading in random geometry. *Eur. Phys. J. B Condens. Matter Complex Syst.* 25, 403–406. doi: 10.1140/epjb/e20020045

Galán, J. M., and Izquierdo, L. R. (2005). Appearances can be deceiving: lessons learned re-implementing axelrod's evolutionary approach to norms. *JASSS* 8.

Giardini, F., and Conte, R. (2012). Gossip for social control in natural and artificial societies. *Simulation* 88, 18–32. doi: 10.1177/0037549711406912

Gilbert, N. (2002). *Varieties of Emergence*. Vol. 32. Chicago, IL: University of Chicago and Argonne National Laboratory.

Gilbert, N. (ed.). (2010). *Computational Social Science*, Volume 1-4 of *SAGE Benchmarks in Social Research Methods*. Thousand Oaks, CA: SAGE Publications Ltd.

Gilbert, N., and Conte, R. (eds.). (1995). *Artificial Societies: The Computer Simulation of Social Life*. Bristol, PA: Taylor & Francis, Inc.

Gilbert, N., and Doran, J. (eds.). (1994). *Simulating Societies. The Computer Simulation of Social Phenomena*. London: UCL Press.

Gilbert, N., and Troitzsch, K. G. (2005). *Simulation for The Social Scientist, 2nd Edn*. Buckingham: Open University Press.

Giles, J. (2011). Social science lines up its biggest challenges. *Nature* 470, 18–19. doi: 10.1038/470018a

Grether, D., Kickhöfer, B., and Nagel, K. (2010). Policy evaluation in multiagent transport simulations. *Transport. Res. Record J. Transport. Res. Board* 2175, 10–18. doi: 10.3141/2175-02

Guerci, E., Rastegar, M., and Cincotti, S. (2010). "Agent-based modeling and simulation of competitive wholesale electricity markets," in *Handbook of Power Systems II, Energy Systems*, eds S. Rebennack, P. M. Pardalos, M. V. F. Pereira, and N. A. Iliadis (Berlin; Heidelberg: Springer), 241–286.

Heckbert, S., Baynes, T., and Reeson, A. (2010). Agent-based modeling in ecological economics. *Ann. N.Y. Acad. Sci.* 1185, 39–53. doi: 10.1111/j.1749-6632.2009.05286.x

Hegselmann, R., and Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis and simulation. *JASSS* 5.

Helbing, D., and Balietti, S. (2011a). From social simulation to integrative system design. *Eur. Phys. J. Spec. Top.* 195, 69–100. doi: 10.1140/epjst/e2011-01402-7

Helbing, D., and Balietti, S. (2011b). "How to do agent-based simulations in the future: from modeling social mechanisms to emergent phenomena and interactive systems design," in *Technical Report 11-06-024* (Santa Fe, NM: Santa Fe Institute).

Helbing, D., Farkas, I., and Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature* 407, 487–490. doi: 10.1038/35035023

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., et al. (2009). Computational social science. *Science* 323, 721–723. doi: 10.1126/science.1167742

Mantegna, R. N., and Stanley, H. E. (2000). *Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge, MA: Cambridge University Press.

Newell, A., and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM* 19, 113–126. doi: 10.1145/360018.360022

Nowak, M. A., and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature* 393, 573–577. doi: 10.1038/31225

Paolucci, M., Kossman, D., Conte, R., Lukowicz, P., Argyrakis, P., Blandford, A., et al. (2013). Towards a living earth simulator. *Eur. Phys. J. Spec. Top.* 214, 77–108. doi: 10.1140/epjst/e2012-01689-8

Pastor-Satorras, R., and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86, 3200–3203. doi: 10.1103/PhysRevLett.86.3200

Pinyol, I., Sabater-Mir, J., Dellunde, P., and Paolucci, M. (2012). Reputation-based decisions for logic-based cognitive agents. *Auton. Agent. Multi-Agent Syst.* 24, 175–216. doi: 10.1007/s10458-010-9149-y

Sakellariou, I., Kefalas, P., and Stamatopoulou, I. (2008). "Enhancing netLogo to simulate BDI communicating agents," in *Artificial Intelligence: Theories, Models and Applications*, Volume 5138 of *Lecture Notes in Computer Science*, eds J. Darzentas, G. Vouros, S. Vosinakis, and A. Arnellos (Berlin; Heidelberg: Springer Berlin Heidelberg), 263–275.

Sawyer, R. K. (2005). *Social Emergence: Societies As Complex Systems*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511734892

Schelling, T. C. (1971). Dynamic models of segregation. *J. Math. Soc.* 1, 143–186. doi: 10.1080/0022250X.1971.9989794

Sornette, D. (2003). *Why Stock Markets Crash Critical Events in Complex Financial Systems*. Princeton, NJ: Princeton University Press.

Windrum, P., Fagiolo, G., and Moneta, A. (2007). Empirical validation of agent-based models: alternatives and prospects. *JASSS* 10.

Wooldridge, M., and Jennings, N. R. (1995). *Intelligent Agents: Theory and Practice*. Available online at: http://www.doc.mmu.ac.uk/STAFF/mike/ker95/ker95-html.h (Hypertext version of Knowledge Engineering Review paper). doi: 10.1007/3-540-58855-8

Xenitidou, M., and Elsenbroich, C. (2010). Construct validity and theoretical embeddedness of agent-based models of normative behaviour. *Int. J. Interdiscip. Soc. Sci.* 5, 67–80. Available online at: http://iji.cgpublisher.com/product/pub.88/prod.1093

Yang, R., Wang, B.-H., Ren, J., Bai, W.-J., Shi, Z.-W., Wang, W.-X., et al. (2007). Epidemic spreading on heterogeneous networks with identical infectivity. *Phys. Lett. A* 364, 189–193. doi: 10.1016/j.physleta.2006.12.021

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 10 June 2014; published online: 14 July 2014.
*Citation: Conte R and Paolucci M (2014) On agent-based modeling and computational social science. Front. Psychol. **5**:668. doi: 10.3389/fpsyg.2014.00668*
This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 02 September 2014
doi: 10.3389/fpsyg.2014.00947

A statistical mechanical problem?

Tommaso Costa*[1]* and Mario Ferraro[2]

[1] *Department of Psychology, University of Turin, Turin, Italy*
[2] *Department of Physics, University of Turin, Turin, Italy*

Edited by:
Giuseppe Boccignone, Università di Milano, Italy

Reviewed by:
Ruben Coen-Cagli, University of Geneva, Switzerland
Rebekka Lencer, University of Muenster, Germany

****Correspondence:***
Tommaso Costa, Department of Psychology, University of Turin, Via Verdi 10, Turin 10124, Italy,
e-mail: tommaso.costa@unito.it

The problem of deriving the processes of perception and cognition or the modes of behavior from states of the brain appears to be unsolvable in view of the huge numbers of elements involved. However, neural activities are not random, nor independent, but constrained to form spatio-temporal patterns, and thanks to these restrictions, which in turn are due to connections among neurons, the problem can at least be approached. The situation is similar to what happens in large physical ensembles, where global behaviors are derived by microscopic properties. Despite the obvious differences between neural and physical systems a statistical mechanics approach is almost inescapable, since dynamics of the brain as a whole are clearly determined by the outputs of single neurons. In this paper it will be shown how, starting from very simple systems, connectivity engenders levels of increasing complexity in the functions of the brain depending on specific constraints. Correspondingly levels of explanations must take into account the fundamental role of constraints and assign at each level proper model structures and variables, that, on one hand, emerge from outputs of the lower levels, and yet are specific, in that they ignore irrelevant details.

Keywords: neural network, statistichal mechanics, behavior, connectivity, mapping

1. INTRODUCTION

Any attempt to derive the processes of perception and cognition or the modes of behavior from sets of neural activities is confronted with the problem of mapping an incredibly large set of possible brain states to a very large number of observables. Simply put, the numbers are staggering: although estimates vary, there are purportedly about $N = 10^{11}$ neurons in the human brain (Sporns, 2012) and even with the very drastic simplification that a neuron is a binary device, possible states are $2^N = 2^{10^{11}}$. This enormous set of states must be mapped into the possible observables and even in this case numbers are huge: for instance even with a conservative estimate the number of possible postures is 10^{30} (Stephens et al., 2011). The sheer orders of magnitude involved seem to prevent the possibility of finding any correspondence among elements of the two sets, i.e., the matching of states to observable processes.

Fortunately there are factors that somewhat simplify the problem: for instance a given behavior can result from many different brain states, as redundancy is a well known evolutionary feature to make living systems more robust. Furthermore brains are made up of very complex networks (connections are of the order of 10^{15}), thus neural states are not independent variables and they tend to form spatio-temporal patterns, rather that disordered sequences of activity. Indeed, fMRI measures have shown that spatial maps of activity are formed even in resting state situations, without any external stimulus (Raichle, 2010). In addition, as suggested in Ganguli and Sompolinsky (2012), states of the dynamical systems describing the activity of cortical areas (e.g., motor cortex, or sensory cortex) are limited by the dimensionality of the inputs (e.g., motor task to be performed, or sensory inputs), which is often much lower than the dimensionality of the cortical dynamical system.

These simplifying factors notwithstanding, the brain is so complex that to explain cognitive and behavioral functions philosophers and scientists have often resorted to conceptual metaphors (Daugman, 1993); modern examples are the computer and information metaphor (see Werner, 2011) for a critical review.

An earlier version of the computation metaphor, based on the seminal work of McCulloch and Pitts (1943), on the equivalence between networks of formal neurons and Turing machines, was centered on the notion that neural activity implements logical calculus via formal rules for the transformation of for the manipulation of symbols (Daugman, 1993), an idea which has provided much impulse to the development of artificial neural networks and their applications (Haykin, 1994; Werner, 2011). The computation metaphor later has given rise to the so called "computational theory" of the brain whose aim is to explain and to simulate the mechanisms by which the brain performs a variety of tasks such as, for instance, edge detection or stereo vision (Marr, 1982). This version of the computation metaphor has became so popular that the term "computational" is nowadays used to characterize almost any model including task analysis (Daugman, 1993).

Complementary to this approach is the information metaphor, that views the brain as an information processing device and focuses on the input–output relations among neurons in the framework of information theory. The central issues in this framework are those of coding and decoding of the neural stimulus, namely which feature of a neural spike train (rate, correlations, etc.) carries the information (in Shannon's sense) and, next, how this information is decoded by the brain, revealing the nature of the external (physical) stimulus (Jacobs et al., 2009; Werner, 2011). The latter problem is known to be an inference problem (Knill and Pouget, 2004), to solve which Bayesian techniques have proven to be very successful. This has lead to the "Bayesian coding hypothesis": the brain represents sensory information in the form of probabilities and derives posterior probabilities of the configurations of the external world (Knill and Pouget, 2004; Doya, 2007; Friston, 2012).

Computation and information metaphors are useful to elucidate important aspects of brain function, but, as pointed out in Werner (2011), they fail to provide the fundamental link between the dynamics of neural activity and computational and information processing properties of the brain. Thus, a different approach has emerged which maintains that real comprehension of cognitive and behavioral functions can only follow from the analysis and explanation of the collective dynamics of neural systems (Werner, 2011; Parker and Srivastava, 2013).

This is also the point of view taken in the present work: specific models related to this approach will be reviewed in more detail later.

Neuronal activity takes place at different scales and a rough classification can distinguish between microscopic (neurons and synapses), mesoscopic (networks and local interactions between neurons), and macroscopic levels (areas of the brain) (Deco et al., 2008). All these levels have their own specificity determined by different types of activity patterns. Then understanding the dynamics of the nervous system requires insights into processes occurring at different scales and that must be matched by appropriate levels of description or representation, characterized by specific variables and model structures.

Different neural models can be represented as elements of a two dimensional space (Cessac and Samuelides, 2007). The first axis of this space describes the type of neuron and its proximity to biology, starting from the Hodgkin–Huxley equations followed by excitable systems with continuous state and finally binary neurons of the McCulloch–Pitts type. The other axis takes in account the collective aspect of neural networks in a hierarchy of ordering: one neuron, few neurons, one population of weakly coupled neurons and finally one population with arbitrary coupling.

Large neural populations present an obvious similarity with physical systems composed of very large number of elements (atoms or molecules) subjected to mutual interactions. In physics the answer to challenges posed by such systems is to resort to mechanical statistical methods, which do not try to solve models at the microscopical level of individual elements, but, instead, use laws of probability to derive a set of collective variables, whose properties can then be studied at the macroscopic level. The success of this approach requires, and indeed depends on, finding the right variables, which can lead to meaningful macroscopic representations, while disregarding irrelevant ones. This, in turn, involves simplifying the system under consideration, from a detailed description to a more abstract representation in which some properties of the elements forming the system are disregarded.

It must be kept in mind, however, there are crucial differences when considering physical vs. neurobiological systems.

1. First, neural systems of the brain are part of living organisms. The problems concerning the transitions from inert to living states of matter and the characterization of life (Smith and Szathmary, 1997; Longo and Montévil, 2012) are outside the scope of this paper. It is enough to say that, at a fundamental level, activity of neural systems is constrained by the amount of metabolic energy available and by the need to limit entropy production (Schrödinger, 1956; Longo and Montévil, 2012).
   More relevant for our work is the fact that animal brains have been shaped by evolutionary pressures and, therefore, neural systems are subjected to many cost-benefits trade-offs, the most basic involving the balance between the speed of respose against the accuracy of identification of a stimulus (Geary, 2005).
   These constraints affect the topology of the connections: empirical evidence suggests that brain anatomical connectivity is locally clustered with a few long-range connections between any pair of regions, and this can be explained by the need to minimize wiring costs while maintaining the possibility of long range interactions among different areas (Bassett and Bullmore, 2006).
2. Neurons interact with the rest of the organism and among themselves in ways, in general, more complex than interactions among elements of physical systems. Furthermore, neurons are computational units, able to perform non trivial computations (Koch, 2004).
3. Differently from physics where the elements of a system can be considered all equal ("all electron are the same" as Fermi put it), neural systems are characterized by heterogeneity, e.g., excitatory vs. inhibitory neurons or electrical vs. chemical coupling.
4. Neural systems are endowed with specific architectures, gauged to specific sensory, motor, and cognitive tasks.
5. Networks can learn by changing the strength of their mutual connections.
6. In physical systems the global behavior can be represented by simple scalars, for instance critical exponents and correlation lengths in non-equilibrium phase transitions, whereas models of large networks in the brain must explain the complex spatio-temporal patterns that make up physiological or behavioral responses. Therefore the question arises of what constitutes the relevant definition of system activity for a given level of explanation.

These differences notwithstanding, a statistical mechanics approach is almost inescapable, since dynamics of the brain as a whole are obviously determined by patterns of neural activities occurring at a lower level, and, indeed statistical mechanics tries to derive the laws at the macroscopic level from interactions among microscopic components.

A classical example are, in the theory of artificial neural networks, the so called Hopfield networks of binary units, (see Hopfield, 1982; Amit, 1992) and, for more recent results, (Advani et al., 2013).

Statistical mechanical techniques are not restricted to the Hopfield model (Coolen and Del Prete, 2003): they have been applied also to biological neural systems both to explain experimental data (Masoller et al., 2009; Montani et al., 2009; Deco et al., 2012) and to provide general models of the brain (Ingber, 1981; Freeman and Vitiello, 2006; Parker and Srivastava, 2013).

It will be argued here that the problem of modeling and representing neural systems of increasing size and complexity is akin to the problem of statistical mechanics and that the way out of the problem of intractability is the same: to assign at each scale proper variables, namely variables that emerge from outputs of the lower level, while ignoring details which are irrelevant for the higher level.

In particular, the main claims of this paper are:

- Systems at each level obey to the laws holding for the lower levels, but they are subjected to new constraints that in turn generate new features, like novel patterns of activity, requiring adequate levels of representations.
- Constraints derive from the neural connections whose complexity increases with the dimension of neural circuits, whose topology then plays a central role in determining neural dynamics.

This approach is inspired by the ideas of Jacob (1977) on the structure of natural systems:

> *"Nature functions by integration.... Each system at a given level uses as ingredients some systems of the simpler level but some only. The hierarchy in the complexity of objects is thus accompanied by a series of restrictions of limitations. At each level new properties may appear that impose new constraint on the system... Those (constraints) that operate at a given levels are still valid at a more complex level."*

2. LEVELS OF COMPLEXITY AND EXPLANATION

Levels of explanations are determined by two main issues: the choice of state variables and the formal structure of the model.

In very general terms, a neural network is a dynamical system describing the temporal evolution of the activities, $\{a_i\}$ $i = 1, \ldots n$, of a neural population of n elements, and can be formally expressed by a map ϕ which starting from the state at initial time t_0 yields the state at time t

$$\phi_i : a_i(t_0) \rightarrow a_i(t) ; \qquad (1)$$

this system can be either deterministic or ruled by probabilistic laws. Maps ϕ_i are usually the solutions of systems of differential equations and their formal expressions are typically very complex, as they depend on a set of external inputs $\{I_j\}$ $j = 1, \ldots m$ and on the connections among neurons. Thus, in general some simplifications are carried out to make the dynamical system more manageable.

First one must decide which variable represents the neural activity: this choice is important not just in order to simplify the problem but because, implicitly, it identifies which aspect of neuron dynamics is considered to be important.

Usually in neural networks theory the elementary computational element is assumed to be the single neuron and the basic variable is the potential V across the membrane, but other, finer, levels of resolution could be considered, for instance ion species and channels or, in principle, the quantum mechanical scale. Suppose, for sake of argument, that it is possible to write down and solve the Schrodinger equation for any molecule or atom of the neuron: the result would be the an incredibly complex wave function which would not explain more than Hodgkin and Huxley theory, because the quantum mechanical scale is not really necessary to understand how spikes are generated, even though, obviously, the laws of quantum mechanics apply to all atoms forming the neuron.

In conclusion, for a neuron a "natural" variable is the difference of potential V across the membrane, whose dynamics are formally described by the theory of Hodgkin and Huxley. However, the level of detail of this model is not really required when one moves from single neurons to neural networks and more abstract models can be developed, whose structure implicitly defines which aspects of spikes generation and transmission are considered important.

For instance, information transmitted along the nervous system of an organism is thought to be encoded by the frequency of the action potentials (the firing rate), and/or by the timing of spikes. Then in modeling the transmission of information one can disregard the shape of the spike and just consider the time intervals with which action potentials occur: this approach is at the basis of the "integrate and fire" type of neuron models (Koch, 2004; Deco et al., 2008).

Neural dynamics can be also described by the temporal variation of spike rates: activity is now identified with the frequency of action potentials and the sequence of spikes collapsed in just one number. The reasons behind this choice, besides the obvious simplification, are based on the observation that many neurobiological phenomena appear to be determined at the level of firing rate. Indeed many experimental data are reported in term of spike rate, which is considered the fundamental element in the information processing in the brain, an idea that goes back to the fundamental work of Adrian (1926). It must be noted that, in recent years, the idea that spike rate suffices to explain coding and decoding of neural signal has been, rather convincingly, questioned (Rieke, 1999).

The rest of this section will try to clarify how increasing complexity of connectivity patterns engenders the emergence of new properties of neural systems and how levels of explanation can be found matching this evolution from simple to complex systems.

2.1. MINIMAL NETWORKS

First we shall consider minimal systems of neuron pairs and a rather abstract and very simple version of the Wilson–Cowan model (Wilson, 1999) will be adopted to illustrate the role of the connections in neural dynamics. The state of the neuron will be represented by its activity a, a real variable, which evolves according to a set of differential equations. It is not important here to give a precise definition of activity, which can be, for instance, V or spike rate, as the equations used in the following can be adapted to different meanings of a. Note that the results described in the sequel are general and not depending on the particular form of the equations, used here solely for illustrative purposes.

The activity of a neuron is described by just one differential equation of the form

$$\tau \frac{da}{dt} = -a + S(I - \theta), \tag{2}$$

where τ is a time constant, θ a threshold and I the total input that can originate from the external world or, more frequently, from other neurons: in this latter case I can be the sum of several inputs. The term $-a$ just expresses the obvious idea that in absence of input the activity relaxes to zero, whereas the function S defines the effect of the input I on the activity a of the neuron and it can be modeled in a variety of ways: usually it is assumed S to be a monotonically increasing function, with $S(I) = 0$ if $I - \theta \leq 0$ and tending to a finite value as I increases.

In the following θ will be set to 0 and for simplicity's sake it will be supposed that, at least in a time interval δt, I is constant. Under these assumptions it is straightforward to show that when the input signal I is switched on the activity a tends to reach a value $a^* = S(I)$. Note that, as S is monotonous, there is an one-to-one relation between a^* and I, so that for any given a^* there exists just one value of I satisfying the equality $a^* = S(I)$; this means that the activity a just scale-transforms the input signal, i.e., reproduces I on a different scale.

Very different, more complex, activity patterns appear in system of mutually connected neurons, even when just two units are considered: the activity is now a vector

$$\mathbf{a} = (a_1, a_2)$$

and the corresponding system can be written as

$$\tau_1 \frac{da_1}{dt} = -a_1 + S\left(w_1 a_2 + I_1\right)$$
$$\tau_2 \frac{da_2}{dt} = -a_2 + S\left(w_2 a_1 + I_2\right) \tag{3}$$

The new elements here are the synaptic weights w_1, w_2, that provide a description of the interaction between the neurons: three cases are possible, each characterized by a specific dynamic:

1. the $\mathcal{E} - \mathcal{E}$ system, where the connections are mutually excitatory, that is $w_i > 0$, $i = 1, 2$,
2. the $\mathcal{I} - \mathcal{I}$ system characterized by mutually inhibitory connections, $w_i < 0$, $i = 1, 2$,
3. the $\mathcal{E} - \mathcal{I}$ system, where one neuron is excitatory and the other is inhibitory, and the synaptic weights have opposite signs.

If the connections are mutually excitatory the network is a bistable system, namely it is characterized by two stable states (attractors): the activity of both neurons can be either low (possibly zero) or it can reach high activity levels, depending on the values of synaptic weights w_1, w_2, and on the inputs I_1, I_2. This very simple network shows that connections between neurons give rise to a set of new behaviors: for instance, also a very short (ideally instantaneous) stimulus to one of the neurons can trigger the evolution of both neurons toward stable high activity levels, i.e., the system is able to self sustain even when the inputs I_1, I_2 are switched off. Attractors of this system are the simplest instance of multi-stabilities, that can be the basis of short time memory (Wilson, 1999) and can provide a mechanism for the switching between different perceptions or behaviors, as suggested by theoretical and experimental studies, (Deco et al., 2007; Moreno-Bote et al., 2007).

Two mutually inhibitory neurons are an elementary example of winner-takes-all networks, which have been widely used in the context of artificial intelligence and pattern recognition. Due to mutual inhibition one of the two neurons has high activity levels whereas the other is not active. The "winning" neuron is determined by the parameters of the system: in particular if $w_1 = w_2$ neuron with the larger input "wins." Such type of network implements the very general principle of competitive exclusion, found also in ecology and population theory, by which when two population compete for resources just one survives (Murray, 2002). Mechanisms of the winner takes all types are thought to be at the basis of selection processes, motor control and path integration (Wilson, 1999).

Finally, a $\mathcal{E} - \mathcal{I}$ system gives rise to the emergence of homoeostasis mechanisms, by which sensory input is regulated, for instance to make the localization of its sources more precise. Moreover the system can be modified in a straightforward way to produce sustained oscillations also in presence of a constant stimulus, an ubiquitous feature in living organisms, from cardiac cycles to the rhythms of breathing and locomotion. Note that this property is unique for the $\mathcal{E} - \mathcal{I}$ arrangement, in that it is straightforward to show with the standard methods of the theory of dynamical systems that such oscillations cannot appear in either mutually excitatory or mutually inhibitory systems. It follows then that oscillations under constant stimulus are due to the heterogeneity of the system, i.e., the presence of both excitatory and inhibitory connections. Pairs of $\mathcal{E} - \mathcal{I}$ type can in turn be connected into coupled oscillators that act as central pattern generators, controlling motion routines (Kleinfeld and Sompolinsky, 1988; Brunel and Wang, 2001).

In conclusion these simple neural systems show that coupling between neurons gives rise to a variety of activity patterns, more complex than those of a single neuron; hence, they exhibit a larger spectrum of computational and behavioral properties. The basis of this enhanced capability resides in the fact that now a_1, a_2 do not depend solely on the inputs, as connections make them dependent one on the other.

Consider a pair of neurons to which are given inputs I_1, I_2, respectively: if they are not connected the attractors of activity are $a_1^* = S(I_1)$, $a_2^* = S(I_2)$. As mentioned before there is a one-to-one correspondence between a^* and I and therefore any activity pair of values a_1^*, a_2^* can be reached given suitable inputs I_1, I_2, since activities are independent one from the other. On the contrary mutual dependence of activities limits the number of states the system can reach. Suppose that neurons now form, for instance, a $\mathcal{I} - \mathcal{I}$ pair. In this case it is straightforward to show that if $w_1 = w_2$ and $I_1 > I_2$ the attractors of this system are $a_1^* = S(I_1)$, $a_2 = 0$, that is the second neuron will be inactive whatever be the value of I_2, provided of course that $I_1 > I_2$.

This idea can be made more precise if one considers activities a_i, $i = 1, 2$, as stochastic variables, with randomness due to fluctuations of the stimulus or to stochasticity in the mechanisms generating the neural response.

The probability that a_i take values in a given interval $a_i + da_i$ can be computed, at least in principle, via the Fokker Planck equation (Deco et al., 2008). The derivation of such equation is not trivial and its solution is, usually, very difficult, but some qualitative results can be readily obtained. Let $p(a_i)$, $i = 1, 2$, be the probability density functions (pdf) of activities a_i and let $p(\mathbf{a})$ be the pdf of the activity vector $\mathbf{a}$; if neurons are supposed to be independent then the entropy $H(\mathbf{a})$ of the stochastic variable $\mathbf{a}$, a measure of disorder of the system, is the sum of the entropies of the single variables a_i, $H(\mathbf{a}) = H(a_1) + H(a_2)$. On the other hand it is a standard result of probability theory that if a_i are not independent, as in case of connected neurons, $H(\mathbf{a}) < H(a_1) + H(a_2)$. Thus, connections among neurons reduce the effect of casual fluctuations and this in turn entails the generation of more complex activity patterns.

Mutual dependence of activities has another important consequence: let the activities a_i, $i = 1, 2$ be the input of some neuron j, and let, for simplicity, assume the weight connecting the input neurons $i = 1, 2$ with j to be equal to 1. The total input reaching j is $I = a_1 + a_2$ and its variance $\sigma_I^2 \leq \sigma_{a_1}^2 + \sigma_{a_2}^2$, where the equality holds only if the activities a_j are independent. We see then that mutual connections provide more reliable global inputs.

2.2. LARGE NEURAL SYSTEMS

It has been shown, so far, that even very simple systems of connected neurons can implement processes of self-organization and entropy reduction. These properties are inherited by large neural populations, but obviously increasing the dimensionality of the system makes the structure of attractors more complex and able to generate a larger number of possible behaviors: for instance more multistabilites appear, that can correspond to a larger number of possible memories or choices. In addition, different experimental techniques (fMRI, EEG, etc.) have shown that the non-linear nature of neural dynamics leads to processes of self- organization and phase transitions (Kelso, 1995; Freeman and Vitiello, 2006).

A variety of theoretical models has been used to investigate the properties of large scale networks: it is not possible here to give a detailed review, but they can be subdivided roughly in models derived by the theory of dynamical systems and models derived by analogies with physical systems.

A natural application of the theory of dynamical systems is the concept of neural field, which represents the organization of the cortex with spatially structured neural network whose dynamics are modeled by differential equations in the continuum limit: activity of neural fields can form dynamic spatio-temporal patterns, similar to the spatial distributions experimentally observed in the brain (Wilson and Cowan, 1972; Deco et al., 2008; Bressloff, 2012).

Other models are derived by analogies with physical systems and use typical methods of statistical mechanics: for instance in Ingber (1981) collective neural activities in the cortex are formulated by considering first the microscopical level of synaptic interactions and averaging them spatially to form a mesoscopical domain. The same procedure is then repeated to produce macroscopic spatial-temporal regions, described by the formalism of stochastic processes. A different, but related, approach (Freeman and Vitiello, 2006) utilizes many-body field dynamics, to derive equations describing ordered pattern formation and phase transitions.

More recently the idea has been put forward that analysis of self organized criticality can provide useful insight in the analysis and function of perceptual, cognitive, and motor networks (Parker and Srivastava, 2013) in that these processes offer a way out from the stability-plasticity dilemma (Abraham and Robins, 2005), namely the opposite requirements of stability and plasticity. Self-organized criticality is a feature of non-linear dynamical systems where the macroscopic behavior of a system emerges from the interactions of its component parts. This results in non-equilibrium phase transitions, i.e., sharp variation of neural activity, which depends on the intrinsic dynamics of the system rather than on external inputs (Parker and Srivastava, 2013).

Neural field and physics based models assume that brain states, and hence behavior, arises from activity propagating from microscopic to mesoscopic and finally to macroscopic scale, and these are the basic levels of explanation.

2.3. THE ROLE OF CONNECTIONS

We have seen that large populations of neurons give rise to a rich variety of behavior. The increased dimensionality of the system leads to the appearance of new topological properties and two principles seem to be at work: segregation and integration. Segregation results in the subdivision of the brain in areas which, for instance, respond to specific sensory inputs or perform specialized tasks.

Next, integration among areas is required to process information coming from different sources in the external world and to produce an appropriate behavioral response (Sporns, 2012).

These dual aspects also provide a structural and functional basis to model brain function. Segregation allows more abstract levels of explanation, in that neurons belonging to the same area can be treated as a single variable, for instance by making use of mean field approximation; any description of the brain activity aiming to explain behavior, however, cannot help but taking into account the topology of the connections at the basis of integration among areas (Geary, 2005; Sporns, 2012).

The increased complexity of neural connections in large populations of neurons suggests that the expansion of the range of possible dynamical states does not depend simply on the number of neurons but also on the more complex interactions occurring within the populations. For instance, consider a set of neurons connected in a purely feedforward way. In this case no oscillatory behavior can arise in response to a constant stimulus, but, on the contrary, a feedback loop may give rise to persistent oscillations.

The relevance of the complexity of connections is apparent in the organization of the visual system, where their topology determines the receptive fields from the retina to simple, complex, and hypercomplex cells in the visual cortex, in that the shapes of these receptive fields require excitatory and inhibitory connections to form a precise configuration (Wandell, 1995). Specific structures also characterize the separate, but interactive, visual systems which preside, respectively, to the formation of internal models of the external world, and to the control of object-directed actions (Goodale and Humphrey, 1998).

Even the generation of simple motions requires specific constraints: the neural population must be subdivided in subnets, each able to perform a specific sub-task, connected in a precise way to ensure an ordered succession of neural events. The simplest gesture requires the coordination of activity of different subnets each firing in a precise sequence: more generally motion routines result from the synchronization of the activity of many oscillators, each composed of several neurons, whose phases must have fixed differences to ensure a proper coordination of single steps (Murray, 2002).

The type and topology of connections appear then to play a crucial role in the functions of brain areas at the macroscopical level. As it could be expected this is true also at the lower scale of activity patterns: for instance spatially localized areas of activity can arise from constant input solely if neurons of the area are linked by mutually excitatory and inhibitory synapses forming the so called "Mexican hat" weights distribution (Murray, 2002; Bressloff, 2012). Also, it has been shown that dominant patterns of spontaneous activities in the brain are determined by neural connectivity (Galán, 2008).

Data from fRMI studies on spontaneous brain activity provide further evidence on the role of neural connections. In recent years several studies found that spatiotemporal activity patterns are both complex and consistent across different subjects at rest (Raichle et al., 2001). This evidence poses the question of their origin, namely whether they are the expression of a common cognitive state or the consequence of the constraints imposed by neural connections (Deco et al., 2009).

Several models have been used to predict experimental patterns of activity using connectivity data derived from neuroanatomy. These models represent brain areas as nodes of a graph whose link were derived from neuronatomical data by application of diffusion tensor imaging (DTI). Dynamics of activity of the nodes (brain areas) are simulated, for each model, by a set of differential equations with different variables: membrane potential (Honey et al., 2007; Ghosh et al., 2008), the mean level of spike rate of a neural population (Deco et al., 2009), the phase synchrony of neural oscillations (Cabral et al., 2013). Finally activity can be modeled by a simplified stochastic spin model (Deco et al., 2012).

For each model the correlation between activity of the pairs of nodes (functional connectivity) has been calculated and compared with the correlation between brain areas. The results show, for all models, similar predictions and good agreement between the experimental and simulated correlations (Cabral et al., 2013).

Two conclusions can be drawn from these analyses. First, observed spatiotemporal activity patterns in resting state can be derived just as a consequence of constraints imposed by neural connections among brain regions. Next, note that the models differ by the type of variables and the only common feature is the connectivity matrix, and yet their results are similar. These results, then, support the idea that connectivity is central in the formation of patterns of activity in the brain.

Such idea is at the core of the Connectome project (Bullmore and Sporns, 2009; Sporns, 2012) that intends to understand the complete details of neural connectivity and to construct a map of the complete structural and functional neural connections *in vivo* (Sporns et al., 2005; Hagmann et al., 2007, 2008).

3. DISCUSSION

It is often said that the human brain is the most complex structure in the known universe, even though how such complexity can be computed is still a open question. In Tononi et al. (1994), complexity is derived by measures of mutual information, but other definitions could be considered, based on the entropy of the states of neural populations (Shiner et al., 1999). In any case the complex nature of the brain reveals itself in the structure of its connections and patterns of activities. These two aspects are inextricably linked: the structure of interactions among elements of a neural population generates patterns of activity of increasing complexity.

If the single neuron can just perform a scale transformation of the inputs, pairs of mutually connected neuron can give rise to a variety activity patterns, characterized by the presence of attractors and sustained oscillations. These patterns result from the constraints that weights impose on activities. Also, we have shown that in large neural systems the processes of integration and segregation of connections give rise to a greater variety of activities of neurons and neuron groups.

As mentioned earlier, in models of biological networks events are usually supposed to occur at three canonical scales, namely: microscopic, mesoscopic, and macroscopic, to which correspond different levels of explanation.

Inside each scale some finer subdivision can be considered. For instance, motifs, small repetitive networks occurring in large neural populations and supposed to be building blocks of larger networks (Sporns and Kötter, 2004; Battaglia et al., 2012) can be thought of as an intermediate level between microscopic and mesoscopic scales. Also networks devoted to specific behavioral or cognitive tasks can provide a link between mesoscopic and macroscopic levels.

An interesting suggestion has been presented in West and Deering (1994): in many physical systems "*exists a critical dimension above which fluctuations have only a quantitative effect, but below which the fluctuation can be amplified to modify the qualitative behavior of the phenomenon.*"

In the context of neurobiology, this observation could be translated to mean that domains in the cortex in which variations of activity are amplified into sharp transitions implicitly determine a proper scale for the explanation, for instance, of the sensory or cognitive responses to an input.

The focus of the present work is on the connectivity among neurons in large neural populations and considers a simple neuron model with complex connections, so it can be thought of as situated close to one end of the conceptual space proposed in Cessac and Samuelides (2007); moving across this space one can find models with different emphasis on the neuron/connectivity relationships. At the opposite end of the spectrum with respect to the approach presented here is the analysis of the computational properties of the single neuron, which appears to be able to perform also complex computations (Rieke, 1999; Dayan and Abbott, 2001; Koch, 2004). Each specific model can be backed (or disproved) by specific types of data, from recording of electrical activity for single neurons or small networks to activity maps, for instance obtained with fMRI techniques, for large populations.

ACKNOWLEDGMENTS

We thank the reviewers for many useful suggestions that have helped to improve a previous version of this work. We thank also Prof. Terry Caelli for an helpful reading of the manuscript.

REFERENCES

Abraham, W. C., and Robins, A. (2005). Memory retention–the synaptic stability versus plasticity dilemma. *Trends Neurosci.* 28, 73–78. doi: 10.1016/j.tins.2004.12.003

Adrian, E. D. (1926). The impulses produced by sensory nerve endings part i. *J. Physiol.* 61, 49–72.

Advani, M., Lahiri, S., and Ganguli, S. (2013). Statistical mechanics of complex neural systems and high dimensional data. *J. Stat. Mech. Theory Exp.* 2013:P03014. doi: 10.1088/1742-5468/2013/03/P03014

Amit, D. J. (1992). *Modeling Brain Function: The World of Attractor Neural Networks.* Cambridge, UK: Cambridge University Press.

Bassett, D. S., and Bullmore, E. (2006). Small-world brain networks. *Neuroscientist* 12, 512–523. doi: 10.1177/1073858406293182

Battaglia, D., Witt, A., Wolf, F., and Geisel, T. (2012). Dynamic effective connectivity of inter-areal brain circuits. *PLoS Comput. Biol.* 8:e1002438. doi: 10.1371/journal.pcbi.1002438

Bressloff, P. C. (2012). Spatiotemporal dynamics of continuum neural fields. *J. Phys. A Math. Theor.* 45:033001. doi: 10.1088/1751-8113/45/3/033001

Brunel, N., and Wang, X. J. (2001). Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J. Comput. Neurosci.* 11, 63–85. doi: 10.1023/A:1011204814320

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2575

Cabral, J., Kringelbach, M. L., and Deco, G. (2013). Exploring the network dynamics underlying brain activity during rest. *Prog. Neurobiol.* 114, 102–131. doi: 10.1016/j.pneurobio.2013.12.005

Cessac, B., and Samuelides, M. (2007). From neuron to neural networks dynamics. *Eur. Phys. J. Spec. Top.* 142, 7–88. doi: 10.1140/epjst/e2007-00058-2

Coolen, A., and Del Prete, V. (2003). Statistical mechanics beyond the hopfield model: solvable problems in neural network theory. *Rev. Neurosci.* 14, 181–194. doi: 10.1515/REVNEURO.2003.14.1-2.181

Daugman, J. G. (1993). "Brain metaphor and brain theory," in *Computational Neuroscience*, ed E. L. Schwartz (Cambridge, MA: MIT Press), 9–18.

Dayan, P., and Abbott, L. F. (2001). *Theoretical Neuroscience*, Vol. 31. Cambridge, MA: MIT press.

Deco, G., Jirsa, V., McIntosh, A., Sporns, O., and Kötter, R. (2009). Key role of coupling, delay, and noise in resting brain fluctuations. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10302–10307. doi: 10.1073/pnas.0901831106

Deco, G., Jirsa, V. K., Robinson, P. A., Breakspear, M., and Friston, K. (2008). The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* 4:e1000092. doi: 10.1371/journal.pcbi.1000092

Deco, G., Scarano, L., and Soto-Faraco, S. (2007). Weber's law in decision making: integrating behavioral data in humans with a neurophysiological model. *J. Neurosci.* 27, 11192–11200. doi: 10.1523/JNEUROSCI.1072-07.2007

Deco, G., Senden, M., and Jirsa, V. (2012). How anatomy shapes dynamics: a semi-analytical study of the brain at rest by a simple spin model. *Front. Comput. Neurosci.* 6:68. doi: 10.3389/fncom.2012.00068

Doya, K. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding.* Cambridge, MA: MIT Press.

Freeman, W. J., and Vitiello, G. (2006). Nonlinear brain dynamics as macroscopic manifestation of underlying many-body field dynamics. *Phys. Life Rev.* 3, 93–118. doi: 10.1016/j.plrev.2006.02.001

Friston, K. (2012). The history of the future of the bayesian brain. *Neuroimage* 62, 1230–1233. doi: 10.1016/j.neuroimage.2011.10.004

Galán, R. F. (2008). On how network architecture determines the dominant patterns of spontaneous neural activity. *PLoS ONE* 3:e2148. doi: 10.1371/journal.pone.0002148

Ganguli, S., and Sompolinsky, H. (2012). Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* 35, 485–508. doi: 10.1146/annurev-neuro-062111-150410

Geary, D. C. (2005). *The Origin of Mind.* Washington, DC: American Psychological Association.

Ghosh, A., Rho, Y., McIntosh, A. R., Kötter, R., and Jirsa, V. K. (2008). Noise during rest enables the exploration of the brain's dynamic repertoire. *PLoS Comput. Biol.* 4:e1000196. doi: 10.1371/journal.pcbi.1000196

Goodale, M. A., and Humphrey, K. G. (1998). The objects of action and perception. *Cognition* 67, 181–207. doi: 10.1016/S0010-0277(98)00017-1

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol.* 6:e159. doi: 10.1371/journal.pbio.0060159

Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V. J., Meuli, R., et al. (2007). Mapping human whole-brain structural networks with diffusion mri. *PLoS ONE* 2:e597. doi: 10.1371/journal.pone.0000597

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation.* Upper Saddle River, NJ: Prentice Hall PTR.

Honey, C. J., Kötter, R., Breakspear, M., and Sporns, O. (2007). Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10240–10245. doi: 10.1073/pnas.0701519104

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Ingber, L. (1981). Towards a unified brain theory. *J. Soc. Biol. Struct.* 4, 211–224. doi: 10.1016/S0140-1750(81)80037-1

Jacob, F. (1977). Evolution and tinkering. *Science* 196, 1161–1166. doi: 10.1126/science.860134

Jacobs, A. L., Fridman, G., Douglas, R. M., Alam, N. M., Latham, P. E., Prusky, G. T., et al. (2009). Ruling out and ruling in neural codes. *Proc. Natl. Acad. Sci. U.S.A.* 106, 5936–5941. doi: 10.1073/pnas.0900573106

Kelso, J. S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior.* Cambridge, MA: MIT press.

Kleinfeld, D., and Sompolinsky, H. (1988). Associative neural network model for the generation of temporal patterns. theory and application to central pattern generators. *Biophys. J.* 54, 1039–1051. doi: 10.1016/S0006-3495(88)83041-8

Knill, D. C., and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Koch, C. (2004). *Biophysics of Computation: Information Processing in Single Neurons.* New York, NY: Oxford University Press.

Longo, G., and Montévil, M. (2012). The inert vs. the living state of matter: extended criticality, time geometry, anti-entropy–an overview. *Front. Physiol.* 3:39. doi: 10.3389/fphys.2012.00039

Marr, D. (1982). *Vision: A Computational Investigation Into The Human Representation and Processing of Visual Information.* New York, NY: Henry Holt and Co Inc.

Masoller, C., Torrent, M., and García-Ojalvo, J. (2009). Dynamics of globally delay-coupled neurons displaying subthreshold oscillations. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 367, 3255–3266. doi: 10.1098/rsta.2009.0096

McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. doi: 10.1007/BF02478259

Montani, F., Ince, R. A., Senatore, R., Arabzadeh, E., Diamond, M. E., and Panzeri, S. (2009). The impact of high-order interactions on the rate of synchronous discharge and information transmission in somatosensory cortex. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 367, 3297–3310. doi: 10.1098/rsta.2009.0082

Moreno-Bote, R., Rinzel, J., and Rubin, N. (2007). Noise-induced alternations in an attractor network model of perceptual bistability. *J. Neurophysiol.* 98, 1125–1139. doi: 10.1152/jn.00116.2007

Murray, J. D. (2002). *Mathematical Biology*, Vol. 2. Berlin; Heidelberg: Springer-Verlag.

Parker, D., and Srivastava, V. (2013). Dynamic systems approaches and levels of analysis in the nervous system. *Front. Physiol.* 4:15. doi: 10.3389/fphys.2013.00015

Raichle, M. E. (2010). The brain's dark energy. *Sci. Am.* 302, 44–49. doi: 10.1038/scientificamerican0310-44

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* 98, 676–682. doi: 10.1073/pnas.98.2.676

Rieke, F. (1999). *Spikes: Exploring The Neural Code.* Cambridge, MA: MIT Press.

Schrödinger, E. (1956). *What is Life?: And Other Scientific Essays*, Vol. 88. Doubleday.

Shiner, J. S., Davison, M., and Landsberg, P. T. (1999). Simple measure for complexity. *Phys. Rev. E* 59:1459. doi: 10.1103/PhysRevE.59.1459

Smith, J. M., and Szathmary, E. (1997). *The Major Transitions in Evolution.* Oxford University Press.

Sporns, O. (2012). *Discovering the Human Connectome.* Cambridge, MA: MIT Press.

Sporns, O., and Kötter, R. (2004). Motifs in brain networks. *PLoS Biol.* 2:e369. doi: 10.1371/journal.pbio.0020369

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042

Stephens, G. J., Osborne, L. C., and Bialek, W. (2011). Searching for simplicity in the analysis of neurons and behavior. *Proc. Natl. Acad. Sci. U.S.A.* 108, 15565–15571. doi: 10.1073/pnas.1010868108

Tononi, G., Sporns, O., and Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. U.S.A.* 91, 5033–5037. doi: 10.1073/pnas.91.11.5033

Wandell, B. A. (1995). *Foundations of Vision.* Sunderland, MA: Sinauer Associates.

Werner, G. (2011). Letting the brain speak for itself. *Front. Physiol.* 2:60. doi: 10.3389/fphys.2011.00060

West, B. J., and Deering, W. (1994). Fractal physiology for physicists: lévy statistics. *Phys. Rep.* 246, 1–100. doi: 10.1016/0370-1573(94)00055-7

Wilson, H. R. (1999). *Spikes, Decisions, and Actions: The Dynamical Foundations of Neuroscience*, Vol. 5. Oxford: Oxford University Press Oxford.

Wilson, H. R., and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12, 1–24. doi: 10.1016/S0006-3495(72)86068-5

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 February 2014; accepted: 07 August 2014; published online: 02 September 2014.

*Citation: Costa T and Ferraro M (2014) A statistical mechanical problem? Front. Psychol. **5**:947. doi: 10.3389/fpsyg.2014.00947*

This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 16 September 2014
doi: 10.3389/fpsyg.2014.01034

The concepts of representation and information in explanatory theories of human behavior

Renato T. Ramos*

Laboratory of Psychophysiology and Neurophysiology (LIM-23), Department of Psychiatry, Institute of Psychiatry, University of São Paulo Medical School, São Paulo, Brazil

Edited by:
Roberto Cordeschi, Sapienza University of Rome, Italy

Reviewed by:
Marcello Frixione, Università di Genova, Italy
Ernesto Burattini, Università degli Studi di Napoli Federico II, Italy

Correspondence:
Renato T. Ramos, Laboratory of Psychophysiology and Neurophysiology (LIM-23), Department of Psychiatry, Institute of Psychiatry, Hospital das Clinicas, University of São Paulo Medical School, Rua Dr. Ovídio Pires de Campos, 785 São Paulo, SP 05403-010, Brazil
e-mail: rtramos@usp.br

Focusing in experimental study of human behavior, this article discusses the concepts of information and mental representation aiming the integration of their biological, computational, and semantic aspects. Assuming that the objective of any communication process is ultimately to modify the receiver's state, the term correlational information is proposed as a measure of how changes occurring in external world correlate with changes occurring inside an individual. Mental representations are conceptualized as a special case of information processing in which correlational information is received, recorded, but also modified by a complex emergent process of associating new elements. In humans, the acquisition of information and creation of mental representations occurs in a two-step process. First, a sufficiently complex brain structure is necessary to establishing internal states capable to co-vary with external events. Second, the validity or meaning of these representations must be gradually achieved by confronting them with the environment. This contextualization can be considered as part of the process of ascribing meaning to information and representations. The hypothesis introduced here is that the sophisticated psychological constructs classically associated with the concept of mental representation are essentially of the same nature of simple interactive behaviors. The capacity of generating elaborated mental phenomena like beliefs and desires emerges gradually during evolution and, in a given individual, by learning and social interaction.

Keywords: information, mental representation, human behavior

INTRODUCTION

The construction of comprehensive explanatory models of human behavior requires constant review and improvement of concepts in order to integrate different types of structures and levels of implementation. In this sense, this article discuss two concepts frequently used for modeling different aspects of human behavior in biological, psychological, philosophical, physical, and computational explanatory theories. They are the concepts of information and representation. The objective is to discuss the interdependency between both constructs with special attention to their use in experimental investigations of cognitive phenomena.

Briefly, the idea of representation discussed here is related to the brain's capacity of developing inner states, in the form of relatively stable patterns of neuronal activity, that keep some kind of relationship with events occurring in external world. In many cases, these representations start by simple reactions to external stimuli but, due to brain's organizational characteristics, evolve by incorporating many other elements than those directly apprehensible from the direct contact with the environment. This capacity of constructing complex mental representations results from a long evolutionary process but its basic constituents can be identified in the neuronal activity of simpler organisms in the form of reactive or conditioned behaviors, for example.

The concept of mental representation in cognitive sciences is frequently associated to complex phenomena such as beliefs and desires. This class of models, also known as representational theories of mind (RTM), consider that these states have "intentionality" in the sense that they are *about* or *refer to* things, and may be evaluated with respect to properties like consistency, truth, appropriateness, and accuracy (Cummins, 1989).

This article proposes that the general idea of intentionality or the propriety of mental states of maintaining a correlation with external events can be generalized to describe even the early stages of information processing in the nervous system. This mechanism of co-variation, in association with memory resources and the capacity of generating brain states related to abstract elements of world (more specifically the capacity of deduce the rules governing the behavior of external elements) allow the emergence of the characteristically human cognitive traits.

This broad idea of intentionality is based in a peculiar concept of information as a linking element between brains and world. Information, as used in neurobiological research, can be described as something intrinsically linked to the construction of representations but at the same time as a concept not exclusive of mental instance. Information seems to exist in natural world and human mind has a very special capability of extracting, processing, and using it to increase its capacity of interaction with the environment.

Although frequently studied separately, the concepts of information and representation can be described as having computational and semantic aspects. The term computational refers to the possibility of codification, quantification, manipulation, and physical implementation of information and representations while the term semantic refers to the meaning of both concepts in different contexts.

Information and representation will be discussed here from a neurobiological point of view but with the intention of maintaining coherence with their conceptualization in computational or artificial models of cognition. This coherence requires considering mental representations as biological phenomena, proper but not exclusive of human minds, which construction is achieved by a mechanism of information exchange with the external world. As we shall see below, although representations can be localized in the brain, their meaning does not reside exclusively in the neurobiological instance being a characteristic of the dynamic interaction between brains and environment.

In the following sections, the concepts of mental representation and information will be discussed with a declared bias toward its application in empirical problems of cognitive neurosciences. The interest in these concepts, however, is not restrict to the study of human cognition. Comprehensive discussions about classic information theory can be found in Shannon (1948), Karnani et al. (2009), Wang and Shen (2011), and Adami (2012). The nature of mental representations in philosophy, psychology, and neurosciences is discussed by Cummins (1989, 1996), Stich (1992), and Fodor (2000). Comprehensive discussions about semantic information are found in Floridi (2005), Karnani et al. (2009), Jensen et al. (2013), and Vakarelov (2014).

THE EMPIRICAL STUDY OF HUMAN BEHAVIOR

The paradigmatic situation faced by neuroscientists during their experimental work can be described as follows: consider an individual observing an object and/or carrying out a mental task while his/her brain activity is recorded by a functional neuroimaging machine. Based in the machine's outputs, the scientist controlling the experimental setup wants to know how the individual's cognition works and to what extent the machine output reflects the individual experience of thinking.

Although it is possible to get some kind of information from the machine, the descriptive capacity of this paradigm is limited, especially in relation to the apprehensibility of subjective experiences. This limitation can be expressed by the qualia argument: although the scientist can know something about the individual's internal state it is impossible for an external observer to have access to the very nature of mental processes because they involve a special quality of conscious experience that cannot be reduced to a linguistically mediated set of descriptive elements (Kanai and Tsuchiya, 2012; Ramos, 2012).

This problem can be partially reduced by questioning the individual about her/his subjective experience and checking the accuracy of her/his representations of the external world. This method, however, is limited by the capacity of the individual in accessing their own internal states. The extra-conscious character of many brain activities makes it impossible for someone to be aware and report all elements composing the cognitive experience. Even simple activities are subject to uncontrollable perceptive distortions (optical illusions, for example), spontaneous evocation of memory contents, and subtle affective states that are not consciously perceived.

Although neuroimaging techniques do not account for the qualia question, they are continually improving their capacity of detecting details of brain function in terms of anatomic location and time course of events. The information obtained by functional imaging machines is expressed in terms of electrical signals or measures of metabolic activity which must be articulated with the individual's linguistic descriptions. Machine recordings allow the spatial localization of structures working at a given moment as well as mapping the time dynamics of their interaction (Nunez et al., 2014). Thus, functional data are collected and analyzed based in a general conceptualization of the brain as an information processing device constituted by specialized and widely interconnected substructures working in constant communication.

INFORMATION BASED ON RECEIVER

Probably, the most influential theory of information is that proposed by Shannon (1948) based in the concept of entropy or the uncertainty associated to the occurrence of a message. The general communication system proposed by Shannon is shown in **Figure 1**.

In a simplified form, this definition is based on the probability of occurrence of a given message among other possible ones. Although widely explored in computer sciences as well as in the study of interactions between neurons and cortical areas (Bezzi, 2007; Ward and Mazaheri, 2008), this approach is not suitable for many other applications in neurosciences. An accurate estimation of the message probability requires previous knowledge of how many other possible messages can possibly occur, which is frequently inaccessible in behavioral studies. In addition, Shannon's model explicitly does not take in consideration the meaning of the message emitted, transmitted, and received.

The question of meaning of information, centrally important in neurosciences, has been discussed under the general topic of semantic information. Despite the lack of consensus about its definition, semantic information can be described as the data and its meaning, including or not the conditions of truthfulness (Vakarelov, 2014). The study of semantic information has focused in a number of problems, most of them systematized by Floridi (2005). The main question related to the semantic aspect of information of particular interest for this discussion is "how can data acquire their meaning" (Floridi, 2005; Vakarelov, 2010). Vakarelov (2010), for example, suggests a "pragmatic approach to information where one defines the notion of information system as a special kind of purposeful system emerging within the underlying dynamics of the world, and defines semantic information as the currency of the system. In this way, systems operating with semantic information can be viewed as patterns in organized systems."

Returning to the general framework of Shannon's communication system, one can says that the information transmission process is not dependent on the meaning of the message only until reaching the receiver component of the communication system. It

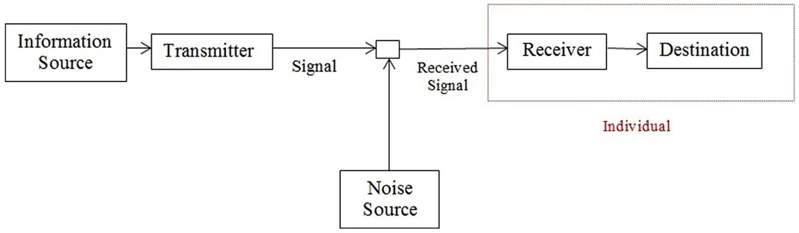


FIGURE 1 | Information system proposed by Shannon. The red square delimits the receiver component of the system where the information's semantic value is determined.

occurs because the objective of sending a message is ultimately to provoke *changes* in receiver's state. These changes are what determine the existence of the message from the receiver's point of view. For example, let's consider an individual in a dark cave populated by bats. In the absence of light and without the capacity of perceiving ultrasounds, the individual can construct only a very partial representation of the cave environment. He/she cannot determine how many bats are inside the cave, what they are doing, and if they are communicating with each other. The observer's state cannot be modified by the events occurring in the cave due to the absence of adequate sensory mechanisms. For the bats, however, the same environment is full of meaningful information due to their capacity of emitting high frequency sounds and analyzing its echoes. If this individual is a scientist interested in understanding bat behavior, he/she can develop instruments to detect ultrasounds otherwise unperceivable and "extract" more information from the environment. Even with this new instrument, the "meaning" of this new information is not immediately clear. The only way to construct a comprehensible picture of bat activities is by establishing correlations between observable behaviors and the signals obtained by the machine. Although it is impossible for the scientist to get full access to the bat's mind and to know how is to be like a bat, he/she can map the modifications observed in the environment and compare them with the modifications occurring in the machine states. If the machine is sufficiently precise and the bat's behavior is sufficiently sophisticated, the scientist can build a limited map of bat's mind.

This example can be extended to the neuroimaging techniques in general. In brain functional studies the strategy of simply correlating stationary brain states with static external stimuli has been proved meaningless. The simple mapping of all neurons firing at the moment that a specific stimulus is presented does not guarantee that the neural activity observed is related to that act of observation. In order to determine the correlation level between external world and internal brain activity, the strategy is to induce changes in object's characteristics and observe the resulting changes occurring in brain activity. In functional brain techniques, co-varying patterns of brain activities and object presentation are usually obtained through several repetitions of stereotyped tasks which results are submitted statistical analysis. In fact, the term stimulus used in biological research can be defined as any modification of the environment that interferes with the organism's state. In this situation, the scientist can check if the observer's brain is receiving information by identifying changes in neural activity that correlate with changes occurring in the external world.

Therefore, the process that defines the information as something significant occurs in the *receiver* component of the system (the red box in **Figure 1**). It does not mean that other components are not relevant but the hypothesis to be explored in the next sections is that the meaning of message emerges in the receiver and any other stimuli running through the information system that is not recognized or that does not induces modifications in the receiver's state is not information.

The Shannon communication system model has been applied in modeling each step of the nervous system's functioning. External stimuli work as an information source to sensory cells that generate action potentials and excite the next neuron in the pathway. Cortical areas work as transmitters and receivers in relation to other areas and one person can also be modeled as transmitter, receiver, noise source, or information media according to the interest of the model. Thus, the limits of each component of an information system in an organism are arbitrary and the same formalism used to describe the interaction between two neurons can, in principle, be applied to describe the interactions between neuron nuclei or even between individuals in social interaction.

DEVELOPING REPRESENTATIONS

The co-variation of an observer's neural/mental states with changes occurring in the external world is the first condition for establishing a representation of objects. Many forms of representation can be generated by this process and several of them may be incomplete or inaccurate. The construction of a set of valid and useful representations requires a complementary mechanism of validation and improvement that, in biological organisms, can be implemented by the process of natural selection.

Tononi (2008) suggests that "through natural selection, epigenesis, and learning, informational relationships in the world mold informational relationships within the main complex that "resonate" best on a commensurate spatial and temporal scale. Moreover, over time these relationships will be shaped by an organism's values, to reflect relevance for survival. This process can be envisioned as the experiential analog of natural selection. As is well known, selective processes act on organisms through differential survival to modify gene frequencies (genotype), which in turn leads to the evolution of certain body forms and behaviors (extrinsic phenotype)."

Thus, the acquisition of information and creation of mental representations occurs in a two-step process. First, a sufficiently complex brain structure is necessary to establishing internal states capable to co-vary with external events. Second, the validity of these representations must be gradually achieved by confronting them with the environment. The hypothesis discussed here is that the sophisticated psychological constructs classically associated with the concept of mental representation start from simple interactive behaviors. The capacity of using language and interacting in social groups allowed the gradual emergence of more complex human mental phenomena. This development can had occurred even by a relatively disorganized process of creation, modification, and correction of internal states in function of new inputs from external world.

Therefore, it is possible to admit that the mechanisms by which human cognition had developed are present in other classes of organisms. For example, an insect survives in its natural habitat because it can maintain a sufficiently accurate representation of external world. This representation-mediated "world-insect relationship" is limited and it even may not be considered as of cognitive nature. However, the quality and precision of this representation is the optimized result of a compromise between anatomo-physiological constraints and the necessity of providing information processing resources in the context of selective pressure in a specific niche. Partial representations may be suited to improve survival chances because they are easier to be created and corrected and faster to be implemented in natural life situations.

REPRESENTING RULES

Another representational strategy that emerged along the evolution is the representation of the rules or patterns governing what happen in the external world. For example, conditioned behaviors in several animal species can be understood as a representation of external regularities. The increased dog's salivation after a conditioned stimulus related to food is mediated by a representation, established by learning, of a rule of correlation between two events.

In the human brain, similar mechanisms seem to work even in more complex activities. Noelle (2012) reviewed evidences that rule-guided behaviors in humans are associated with the functioning of the prefrontal cortex, the basal ganglia, and related brain structures. The author suggests that a "dopamine-based gating mechanism interacts with standard models of synaptic plasticity to support the development of appropriately isolated and dimensional prefrontal representations, giving rise to improved generalization to novel situations when adequately diverse training experiences are provided." According to this proposal, some regions of the prefrontal cortex may encode references or "pointers" to other prefrontal areas in a representational scheme that would allow for essentially combinatoric generalization to novel rules. This capacity of combinatoric generalization does not imply a "mere implementation" of symbolic rule-interpretation mechanisms. For Noelle, "complex interactions between the rule representations actively maintained in prefrontal cortex and the dynamic processes of more posterior neural circuits give rise to graded and context-sensitive patterns of performance that escape description by a purely symbolic rule account. Also, statistical regularities in the experiences present during the development of prefrontal cortex can profoundly shape the kinds the explicit rules that can robustly be represented and applied."

The process of information processing based on representation of rules can be further enhanced by the creation of subsets of *a priori* representations available for use in natural situations. Innate behaviors, related to threat detection for example, require the pre-existence of relatively complex representations capable of enhancing fast protective actions. This characteristic is called preparedness of fear and phobias and it has been identified also in human behavior. Mineka and Ohman (2002) present evidences for the existence of an evolved module for fear elicitation and fear learning with four primary characteristics: "First, it is preferentially activated by stimuli related to survival threats in evolutionary history. Thus, fear-relevant stimuli lead to superior conditioning of aversive associations compared with fear-irrelevant stimuli. Second, the module is automatically activated by fear-relevant stimuli, meaning that fear activation occurs before conscious cognitive analysis of the stimulus can occur. Third, the fear module is relatively impenetrable to conscious cognitive control, and fear conditioning with fear-relevant stimuli can occur even with subliminal conditioned stimuli. Fourth, the amygdala seems to be the central brain area dedicated to the fear module."

The high velocity required by the process of identifying threats and implementing adequate responses imply in an increased probability of errors related to the simplification of external situations, misinterpretation of new events, and ultimately the creation of distorted representations. This style of cognitive functioning can be understood under a biological perspective where, in natural situations, errors of commission (wrongly reacting to a non-threat) are more acceptable than errors of omission (not reacting to a real threat).

Other cognitive capacities like empathy and face recognizing also seem to be implemented by similar mechanisms of working with pre-prepared representations (Regenbogen et al., 2012; Kryklywy et al., 2013; Prochnow et al., 2013). Admitting that the same design strategy is used in the implementation of other cognitive functions, this mechanism of simplifying representations in order to facilitate stimuli responses may be hypothesized as playing a role in complex phenomena associated to partial or biased evaluations of external situations like folk psychological explanations and the occurrence of preconceptions in social contexts.

CORRELATION AND INFORMATION

In order to differentiate from Shannon's informational entropy, the term *correlational information* is proposed here, not as a measure of probability but as a measure of how changes occurring in external world correlate with changes occurring inside an agent. This concept does not depend either on the physical, biological, or linguistic nature of external object nor on the cognitive capacity of the receiver. Correlational information depends on the receiver capacity of modifying aspects of its internal states in function of changes occurring in the external environment. This receiver's plasticity needs not to reflect every characteristic of external objects because even partial representations can be

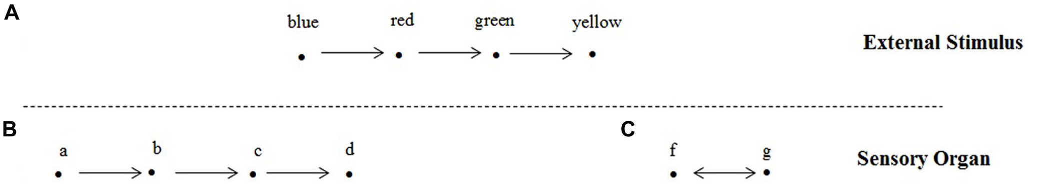


FIGURE 2 | This figure illustrates different representational capacities of sensory organs. (A) represents a light source emitting different colors. The sensory organ illustrated in **(B)** is capable of associating the series of different states a,b,c,d where each state is related to a different color blue, red, green, and yellow. **(C)** illustrate another kind of sensory organ not capable of distinguishing blue from green and red from yellow and, therefore, representing the changes occurring in the exterior world by simplified set of states (f,g).

sufficient for adequate interactions with the environment. This strategy of adopting an information model based in correlations aims to emphasize the importance of the receiver component of in the general model of information system.

This approach can be illustrated as follow: let's consider an animal visually perceiving a light source emitting different colors (**Figure 2A**). If its sensory organ has the capacity of having its state modified in a given way by each color which induces one corresponding change of state (**Figure 2B**), one can says that this animal is capable of having accurate color perception. Note that, in this model, how exactly this correspondence is physically implemented is not important. The central point is that the path blue, red, green, yellow in the external world correspond to the path a,b,c,d inside the organism. In contrast, if the sensory organ is not capable of distinguishing blue from green and red from yellow, for example, its internal representation is given by a simpler path (fig) in **Figure 2C**.

The representation expressed in **Figure 2C** is partial in comparison to that expressed by **Figure 2B** but its physical implementation by a simpler sensory organ demands less resources. If both representations have the same efficiency in preserving the animal's life (detecting food or predators, for example), the simplest alternative may be the most advantageous unless new changes occur in the environment making the exact color perception an essential trait for survival.

According this model, the flux of correlational information along nervous system is the set of modifications gradually established along sensory cells, nerves, interneurons, and brain structures involved in behavior expression. An advantage of this concept is that these modifications are potentially detectable by functional techniques although not always accessible to an individual's consciousness. In experimental context, even physiological manifestations like, for example changes in autonomic functioning or postural control can be considered as part of the set of information that composes mental representations. The inclusion of these not purely cognitive elements is essential, for example, in the study of emotions where several experiential elements cannot be adequately described by language.

This proposal does not imply in denying the existence of internally generated states. Although mental events can occur with a degree of independence from external influences (for example, reflections, interpretations, and mathematical thinking) the basic neural components that allowed the development of these sophisticated capacities are closely related to those working in other relatively more simple brain activities.

The human thinking process can run with a relative independence from external inputs like in mental fantasies. The correlational model proposes is that the ability of working at this level of abstraction was acquired by the gradual improvement of the capacity of using correlational information. Once acquired, this ability allows to the individual to work with independence from direct sensorial inputs and add new elements to mental contents. Although fantasies can be generated with large degree of freedom, the awareness that these contents are internally created is given by the capacity of confronting them with external inputs.

One example of internally generated state involving pre-prepared structures closely related with external events is the mirror neurons system. Originally found in macaque monkeys, in the ventral premotor cortex, area F5 and inferior parietal lobule, this group of neurons fire when the animal sees another animal (or the experimenter) performing actions similar to those pertaining to its natural repertoire of actions. Neuroimaging and electrophysiological studies indicate that mirror neurons may serve for action recognition in monkeys as well as humans, whereas their putative role in imitation and language may be realized in human but not in monkey (Oztop et al., 2013). Although primarily of motor nature, mirror neurons have been associated with mental activities like intention understanding, emotions, empathy, and speech (Acharya and Shukla, 2012).

Another examples of mental representations based in brain-environment co-variant proprieties are those involved in the orientation and movement in the space. Land (2014) points out "that the motor system requires a representation of space that maintains a consistent relationship with objects in the outside world as the body moves within it, then this could also serve as a model of a stable outside world of which we can be conscious. A high-definition representation is not necessary, all that is required is that it provides a stable framework to which detailed information, provided by the visual pathways through the occipital and temporal lobes, can be temporarily attached."

The creation and recording of mental representations involves the gradual recruiting of relatively distant but highly connected brain components with different time dynamics. Consequently, mental representations are not localized in specific brain regions but they gradually emerge along the entire neuronal processing. This idea is compatible with several neurobiological phenomena associated with conscious experience. Shen et al. (2013) proposed that the experience of "insight," described as an experience related to a state of understanding, which emerges into one's conscious awareness with sudden abruptness, involves many distributed brain regions, including the lateral prefrontal cortex, cingulate cortex, hippocampus, superior temporal gyrus, fusiform gyrus, precuneus, cuneus, insula, cerebellum, and some areas of the parietal cortex.

The ability of processing complex concepts and rules governing external events is essential to the emergence of another property of human cognitive systems that is the possibility of anticipating future events. The capacity of preview the occurrence of a given stimulus can be identified even in simple organisms exhibiting conditioned behaviors. For example, the technique of olfactory conditioning of the sting extension response has been extensively used to yield new insights into the rules and mechanisms of aversive learning in insects (Tedjakumala and Giurfa, 2013).

This simple capacity of representing rules can be improved by the development of more complex neural resources. In fact, this capacity vary from one species to other (Seed et al., 2012) and along the cognitive development of each individual (Wellman et al., 2001). Moreover, there are also evidences that this representational capacity do not depend of neuronal mechanisms but also of adequate social and cultural influences (Moriguchi, 2014).

EMERGENCE AND COMPLEXITY

The next question, central for this discussion, is how simple mechanisms of correlation allow the emergence of complex abstractions in the human mind. A possible strategy for clarifying this point is to explore complex systems theories and its applicability at the several structural and organizational levels evolved in the genesis of human behavior.

The idea that complex patterns can spontaneously emerge from simpler components is largely discussed in natural sciences and a number of theoretical ideas have been proposed to explain their occurrence like, for example agent-based models and genetic algorithms (Caticha and Vicente, 2011; Gros, 2013).

One of these theoretical models in particular, known as self-organized criticality (SOC), has received great attention as a possible explanation for the spontaneous emergence of complex patterns both at neural and behavioral levels. The concept of SOC was proposed by Bak et al. (1987) as one of the mechanisms by which complexity arises in nature. They suggested that "dissipative dynamical systems with extended degrees of freedom can evolve toward a self-organized critical state, with spatial and temporal power-law scaling behavior." This spatial scaling leads to self-similar "fractal" structure identifiable in many conditions.

Beggs and Plenz (2003) reported evidences of this phenomenon studying organotypic cultures from coronal slices of rat somatosensory cortex. They continuously recorded spontaneous local field potentials (LFPs) using a 60 channel multielectrode array and found that the propagation of synchronized LFPs activity was described by a power-law. The authors suggested the slope of this power-law, as well as its branching parameter, indicate the presence of SOC in these preparations. (Beggs and Plenz, 2003) found evidence that the critical branching process optimizes information transmission while preserving stability in cortical networks. Simulations showed that a branching parameter at value found in the experimental preparation optimizes information transmission in feed forward networks, while preventing runaway network excitation. The authors called this pattern "neuronal avalanches" and hypothesized that it could be a generic property of cortical networks and represent a mode of activity differing from oscillatory, synchronized, or wave-like network states.

Compatible with the ideas discussed here, the identification of such patterns of functioning seems to depend on the brain functioning in context. El Boustani et al. (2009) studied intracellular activity of 15 neurons in the primary visual cortex of the anesthetized and paralyzed cat. Each neuron was recorded while presenting four full field stimuli through the dominant eye: a drifting grating at the cell's optimal orientation and spatial frequency, a high spatial definition dense noise, a natural image animated with a simulated eye movement sequence, and a grating animated with the same eye movement sequence. The authors found the recordings displayed power-law frequency scaling at high frequencies, with a fractional exponent dependent on the spatio-temporal statistics of the visual stimuli. They also reported that this effect was reproduced in computational models of a recurrent network. They noted "that the power-law relations found here depend on the stimulus, which means that the frequency scaling exponent does not represent a unique signature of cortical network activity, but rather reflects a measure of the dynamic interplay between the sensory evoked activity and the ongoing recurrent network activity."

The possibility of SOC being relevant for explaining complex human behavior was explored by Ramos et al. (2011) who evaluated groups of individuals with and without mental disorders in social interaction during several weeks. Although the behavior of each individual had been very different from other participants in absolute terms, the statistical description of the different groups (individuals with depression, psychosis, mania, and normal controls) showed identical patterns of behavioral variation. In all groups, comparing the behavior of individuals with themselves, small changes of behavior were very frequent while large variations were rare. The characteristic of having the same variation pattern reproduced at different levels of human activity, suggests the presence of self-similarity (Serrano et al., 2008). The curves describing the behavior of all clinical groups and controls showed the same aspect and fitted a power-law. The authors suggested that the presence of self-similarity and power-laws is compatible with the hypothesis that humans in social interaction constitute a system exhibiting SOC.

Self-organized criticality is certainly a promising concept for integrating biological and behavioral aspects of human behavior under the same causal mechanisms but it doubtless requires more empirical investigations (Hidalgo et al., 2014).

A BRIEF COMMENT ON THE SEMANTIC QUESTION

The last important point to be discussed here is the question of the ascription of *meaning* in informational models of cognition. This is a very problematic discussion in the literature that cannot be adequately addressed in the limited scope of this article. However, the empirical research in neurosciences demands some strategy for dealing with this problem due to the impossibility of understanding many aspects of human behavior without considering some form of justification.

A possible provisional strategy is to leave the concept of meaning momentarily aside and explore a utilitarian approach of the mental representations. In a biological perspective, the immediate utility of behaviors and mental representations is increasing individual's survival chances in different contexts. So, although informations and representations have been defined, in this correlational approach, in function of effects observed in the receiver component, their utilitarian character must be apprehended only in the context of the entire communication system.

Naturally, the idea that human cognition was molded by evolutionary mechanisms is not new. Tononi (2008) explain this hypothesis: "Brain mechanisms, including those inside the main complex, are what they are by virtue of along evolutionary history, individual development, and learning. Evolutionary history leads to the establishment of certain species-specific traits encoded in the genome, including brains and means to interact with the environment. Development and epigenetic processes lead to an appropriate scaffold of anatomical connections. Experience then refines neural connectivity in an ongoing manner though plastic processes, leading to the idiosyncrasies of the individual "connectome" and the memories it embeds."

The general concepts of evolution theory have been used for the explanation of several kinds of behaviors and cognitive phenomena. However, this explanatory strategy still needs to be better incorporated by empirical studies. The same attention dedicated to developing neurofunctional techniques must also be dedicated to the identification and analysis to the characteristics of the environment where the behaviors are manifested. For example, this utilitarian characteristic of informational models suggests that future developments in functional brain studies must consider the use of immersive virtual reality setups as a way of controlling the behavioral context.

CONCLUDING REMARKS

This article aimed to address some questions about the use of the representation and information concepts in the context of experimental research in cognitive sciences. The focus in the "information based on the receiver" proposed here is justified by the interest of developing objective approaches to the study of human behavior in biological and semantic terms. This search for new conceptual approaches took the risk of being superficial in its formalism but it was proposed as a first step for the description of the different elements that contribute to the construction of mental representations.

The correlational information concept discussed here aimed to be sufficiently simple to allow a naturalization of the information concept in the sense that all interaction between physical entities can be seen as an informational phenomenon. In this model, the construction of mental representations can be seen as a special case of information processing in which correlational information is received, recorded, but also modified by a complex, emergent, and possibly stochastic process of associating new elements. The validity of these new internally generated constituent elements is granted by its continuous confrontation with new external inputs and by the selection of the most adequate representations in relation to its capacity of improving survival chances.

The hypothesis is that this basic mechanism works in all animal species but, with the improved human brain capacity, it leads to the emergence of higher order or abstract descriptive elements of external objects that allow the prediction of future events. This process is possible by the manipulation of internal states representing not only objects but also the rules governing their behavior. In this model, although the content of correlational information depend on the receiver capacity of creating internal states capable to co-vary with external events, the utility of a given information can be apprehended only by the observation of the entire communication system.

The continuous process of collecting information, creating representations, generating predictions, comparing with outcomes, and adjusting them in order to optimize their accuracy is compatible with several psychological models of learning and cognitive development. This mechanism of correlational representations is also compatible with a Bayesian conception of cognitive functioning where partial or provisional representations work as estimators of *a priori* probabilities in dealing with future events (Tenenbaum et al., 2011).

The ideas discussed here represent a first approach and naturally demand deeper investigations in relation to its theoretical and empirical implications. In theoretical terms, although theories like SOC are promising in explaining human behavior, other mathematical models also deserve attention. Caticha and Vicente (2011), for example, argue that statistical mechanics can leads to aggregated predictions which can be tested against extensive data sets with partial information about populations. The process of exchanging information and learning patterns involved in these models can elicits collective emergent properties not found in individual behaviors.

In relation to the empirical research, this discussion suggests that the integrative study of the computational and semantic elements that compose human experiences will demand significant technical and theoretical improvements. Technically, the combined register of different variables like cortical electric activity, mapping of eye movements, measures of skin galvanic conductance, and postural control obtained during carefully planned cognitive activities emulated in virtual reality environments, for example, can potentially give a deeper comprehension of the mental, affective, and motor events occurring in realistic contexts.

ACKNOWLEDGMENT

The author is grateful to J. T. S. Rebouças and Ronald C. Ranvaud for insightful discussions and review of early paper's versions.

REFERENCES

Acharya, S., and Shukla, S. (2012). Mirror neurons: enigma of the metaphysical modular brain. *J. Nat. Sci. Biol. Med.* 3, 118–124. doi: 10.4103/0976-9668.101878

Adami, C. (2012). The use of information theory in evolutionary biology. *Ann. N. Y. Acad. Sci.* 1256, 49–56. doi: 10.1111/j.1749-6632.2011.06422.x

Bak, P., Tang, C., and Wiesenfeld, K. (1987). Self-organized criticality: an explanation of the 1/f noise. *Phys. Rev. Lett.* 59, 381–384. doi: 10.1103/PhysRevLett.59.381

Beggs, J. M., and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *J. Neurosci.* 23, 11167–11177.

Bezzi, M. (2007). Quantifying the information transmitted in a single stimulus. *Biosystems* 89, 4–9. doi: 10.1016/j.biosystems.2006.04.009

Caticha, N., and Vicente, R. (2011). Agent-based social psychology: from neurocognitive processes to social data F. *Adv. Complex Syst.* 14, 711–731. doi: 10.1142/S0219525911003190

Cummins, R. (1989). *Meaning and Mental Representation*, Cambridge, MA: MIT Press.

Cummins, R. (1996). *Representations, Targets, and Attitudes.* Cambridge, MA: MIT Press.

El Boustani, S., Marre, O., Béhuret, S., Baudot, P., Yger, P., Bal, T., et al. (2009). Network-state modulation of power-law frequency-scaling in visual cortical neurons. *PLoS Comput. Biol.* 5:e1000519. doi: 10.1371/journal.pcbi.1000519

Floridi, L. (2005). Is semantic information meaningful data? *Philos. Phenomenol. Res.* 70, 351–370. doi: 10.1111/j.1933-1592.2005.tb00531.x

Fodor, J. A. (2000). *The Mind Doesn't Work That Way : The Scope and Limits of Computational Psychology*, Cambridge, MA: MIT Press.

Gros, C. (2013). *Complex and Adaptive Dynamical Systems : A Primer*, New York: Springer. doi: 10.1007/978-3-642-36586-7

Hidalgo, J., Grilli, J., Suweis, S., Muñoz, M. A., Banavar, J. R., and Maritan, A. (2014). Information-based fitness and the emergence of criticality in living systems. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10095–10100 doi: 10.1073/pnas.1319166111

Jensen, G., Ward, R., and Balsam, P. (2013). Information: theory, brain, and behavior. *J. Exp. Anal. Behav.* 100, 408–431. doi: 10.1002/jeab.49

Kanai, R., and Tsuchiya, N. (2012). Qualia. *Curr. Biol.* 22, R392–R396. doi: 10.1016/j.cub.2012.03.033

Karnani, M., Paakkonen, K., and Annila, A. (2009). The physical character of information. *Proc. Roy. Soc. Math. Phys. Eng. Sci.* 465, 2155–2175. doi: 10.1098/rspa.2009.0063

Kryklywy, J. H., Nantes, S. G., and Mitchell, D. G. (2013). The amygdala encodes level of perceived fear but not emotional ambiguity in visual scenes. *Behav. Brain Res.* 252, 396–404. doi: 10.1016/j.bbr.2013.06.010

Land, M. F. (2014). Do we have an internal model of the outside world? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369, 20130045. doi: 10.1098/rstb.2013.0045

Mineka, S., and Ohman, A. (2002). Phobias and preparedness: the selective, automatic, and encapsulated nature of fear. *Biol. Psychiatry* 52, 927–937. doi: 10.1016/S0006-3223(02)01669-4

Moriguchi, Y. (2014). The early development of executive function and its relation to social interaction: a brief review. *Front. Psychol.* 5:388. doi: 10.3389/fpsyg.2014.00388

Noelle, D. C. (2012). On the neural basis of rule-guided behavior. *J. Integr. Neurosci.* 11, 453–475. doi: 10.1142/S021963521250029X

Nunez, P. L., Srinivasan, R., and Fields, R. D. (2014). EEG functional connectivity, axon delays and white matter disease. *Clin. Neurophysiol.* doi: 10.1016/j.clinph.2014.04.003 [Epub ahead of print].

Oztop, E., Kawato, M., and Arbib, M. A. (2013). Mirror neurons: functions, mechanisms and models. *Neurosci. Lett.* 540, 43–55. doi: 10.1016/j.neulet.2012.10.005

Prochnow, D., Kossack, H., Brunheim, S., Müller, K., Wittsack, H. J., Markowitsch, H. J., et al. (2013). Processing of subliminal facial expressions of emotion: a behavioral and fMRI study. *Soc. Neurosci.* 8, 448–461. doi: 10.1080/17470919.2013.812536

Ramos, R. T. (2012). The conceptual limits of neuroimaging in psychiatric diagnosis. *AJOB Neurosci.* 3, 52–53. doi: 10.1080/21507740.2012.721856

Ramos, R. T., Sassi, R. B., and Piqueira, J. R. C. (2011). Self-organized criticality and the predictability of human behavior. *New Ideas Psychol.* 29, 38–48. doi: 10.1016/j.newideapsych.2009.12.001

Regenbogen, C., Schneider, D. A., Finkelmeyer, A., Kohn, N., Derntl, B., Kellermann, T., et al. (2012). The differential contribution of facial expressions, prosody, and speech content to empathy. *Cogn. Emot.* 26, 995–1014. doi: 10.1080/02699931.2011.631296

Seed, A., Seddon, E., Greene, B., and Call, J. (2012). Chimpanzee 'folk physics': bringing failures into focus. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 2743–2752. doi: 10.1098/rstb.2012.0222

Serrano, M. A., Krioukov, D., and Boguñá, M. (2008). Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.* 100, 078701. doi: 10.1103/PhysRevLett.100.078701

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shen, W. B., Luo, J. C. L., and Yuan, Y. (2013). New advances in the neural correlates of insight: a decade in review of the insightful brain. *Chin. Sci. Bull.* 58, 1497–1511. doi: 10.1007/s11434-012-5565-5

Stich, S. (1992). What is a theory of mental representation? *Mind* 101, 243–261. doi: 10.1093/mind/101.402.243

Tedjakumala, S. R., and Giurfa, M. (2013). Rules and mechanisms of punishment learning in honey bees: the aversive conditioning of the sting extension response. *J. Exp. Biol.* 216, 2985–2997. doi: 10.1242/jeb.086629

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science* 331, 1279–1285. doi: 10.1126/science.1192788

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242.

Vakarelov, O. (2010). Pre-cognitive semantic information. *Know. Techn. Pol.* 23, 193–226. doi: 10.1007/s12130-010-9109-5

Vakarelov, O. (2014). From interface to correspondence: recovering classical representations in a pragmatic theory of semantic information. *Minds Mach.* 24, 327–351. doi: 10.1007/s11023-013-9318-2

Wang, C., and Shen, H. (2011). Information theory in scientific visualization. *Entropy* 13, 254–273. doi: 10.3390/e13010254

Ward, B., and Mazaheri, Y. (2008). Information transfer rate in fMRI experiments measured using mutual information theory. *J. Neurosci. Methods* 167, 22–30. doi: 10.1016/j.jneumeth.2007.06.027

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* 72, 655–684. doi: 10.1111/1467-8624.00304

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 January 2014; accepted: 29 August 2014; published online: 16 September 2014.

Citation: Ramos RT (2014) The concepts of representation and information in explanatory theories of human behavior. Front. Psychol. ***5****:1034. doi: 10.3389/fpsyg.2014.01034*

This article was submitted to Theoretical and Philosophical Psychology, a section of the journal Frontiers in Psychology.

ADVANTAGES OF PUBLISHING IN FRONTIERS

FAST PUBLICATION

Average 90 days
from submission
to publication

COLLABORATIVE PEER-REVIEW

Designed to be rigorous –
yet also collaborative, fair and
constructive

RESEARCH NETWORK

Our network
increases readership
for your article

OPEN ACCESS

Articles are free to read,
for greatest visibility

TRANSPARENT

Editors and reviewers
acknowledged by name
on published articles

GLOBAL SPREAD

Six million monthly
page views worldwide

COPYRIGHT TO AUTHORS

No limit to
article distribution
and re-use

IMPACT METRICS

Advanced metrics
track your
article's impact

SUPPORT

By our Swiss-based
editorial team

frontiers
EPFL Innovation Park · Building I · 1015 Lausanne · Switzerland
T +41 21 510 17 00 · info@frontiersin.org · frontiersin.org